J. DIFFERENTIAL GEOMETRY
24 (1986) 221–263

SUB-RIEMANNIAN GEOMETRY

ROBERT S. STRICHARTZ

1. Introduction

By *sub-Riemannian geometry* we mean the study of a smooth manifold M equipped with a smoothly varying positive definite quadratic form on a subbundle S (distribution) of the tangent bundle TM, where S is assumed to be *bracket generating* (sections of S together with all brackets generate TM as a module over the functions on M), and the resulting geometric structures that arise in analogy with Riemannian geometry. This is a subject which has been studied by a number of different investigators, more or less independently, from a number of different viewpoints under a number of different names (*singular Riemannian geometry* and *Carnot-Carathéorody metric* are most commonly used).

In this paper we attempt to give a coherent introduction to the subject, taking the point of view that the subject is a variant of Riemannian geometry. The main topic is the study of geodesics. The quantitative structure of a sub-Riemannian manifold is easily seen to be equivalent to giving a contravariant metric tensor field ($g^{jk}(x)$ in local coordinates) mapping T^*M to TM which is nonnegative definite and has a null-space N equal to the annihilator of S in T^*M. We call this a *sub-Riemannian metric*, and in terms of it we can define the length of any piecewise smooth curve which is tangent to S (we will call such curves *lengthy*). By a well-known theorem of Chow (see also Carathéodory [5]) it is possible to connect any two points by such a curve if the manifold is connected (which we will always assume, for simplicity), and so we can endow M with a metric d defined to be the infimum of the lengths of all lengthy curves joining two points. On the other hand, from the sub-Riemannian metric we can write down a system of Hamilton-Jacobi equations on T^*M, and any solution is called a *geodesic*. The two most basic equations

Received October 28, 1985 and, in revised form, June 27, 1986. Research was supported in part by National Science Foundation Grant DMS-8401354.

are: (1) Is every length minimizing curve a geodesic? (2) Is every geodesic locally a length minimizing curve? The answer to the first question is always yes, and it is a consequence of the fundamental theorem of Pontryagin in the calculus of variations. This fact has been stated before without much justification. For the sake of completeness we give a derivation of it in §6. (Unfortunately, the otherwise excellent paper of Gaveau [14] erroneously claims to give a counterexample. The analysis of the same example in Brockett [3] correctly explains the situation, and in [4] he points out the error in Gaveau's paper. Alas, Gaveau's erroneous claim has been repeated in too many of the references! The author is grateful to A. Sanchez-Calle for bringing up this issue.) The second question we are able to answer affirmatively only under an additional hypothesis, which we call the *strong bracket generating hypothesis* (for any nonzero section X of S, TM is generated by S and $[XS]$). This result is proved in §5, and constitutes the main technical contribution of this paper. It is based on a careful study of the exponential map. In contrast to the Riemannian case, the exponential map is never a local diffeomorphism at the origin. In fact $\exp_p: T_p^* \to M$ annihilates the whole subspace N_p. Nevertheless, away from N_p and near the origin in T_p^* the exponential map is a local diffeomorphism under the strong bracket generating hypothesis. This fact is not true in general, so any attempt to answer question (2) more generally will require a new approach.

Once these basic questions are dealt with, it is relatively easy to deal with other matters. In §7 we discuss completeness and prove the analogue of the Hopf-Rinow theorem. In §8 we discuss isometries. There are naturally several notions here, isometry with respect to the metric, infinitesimal isometry (the derivative preserves the sub-Riemannian metric), and regular infinitesimal isometry (the mapping factors through the exponential mapping); it is this last notion which is most useful. We conjecture that every isometry is automatically a regular infinitesimal isometry, but we have not been able to prove this, even assuming the strong bracket generating hypothesis. Our main result is that under this hypothesis the regular infinitesimal isometries form a Lie group and the isotropy subgroup of a point is compact, isomorphic to a compact subgroup of the orthogonal group $O(m)$ (where m is the fiber dimension of S).

In §9 we introduce the notion of a *sub-Riemannian symmetric space*, a space with a transitive Lie group of isometries and a symmetry at each point (however this is not geodesic reflection). On the infinitesimal level, a sub-Riemannian symmetric space is characterized by the following data: a Lie algebra $\mathfrak{g}$, an involutive automorphism which splits $\mathfrak{g} = \mathfrak{g}^+ \oplus \mathfrak{g}^-$, a subalgebra $\mathfrak{k} \subseteq \mathfrak{g}^+$, a subspace $\mathfrak{p} \subset \mathfrak{g}^-$, and a positive definite quadratic form Q on $\mathfrak{p}$,

where $\mathfrak{g}$ is generated by $\mathfrak{k} + \mathfrak{p}$ as a Lie algebra and ad $\mathfrak{k}$ preserves $\mathfrak{p}$ and Q. The space is G/K for suitable Lie groups G and K (compact) with Lie algebras $\mathfrak{g}$ and $\mathfrak{k}$. In §10 we classify locally all three-dimensional examples; they fall into six classes which include Lie groups of semisimple, nilpotent, and solvable type. This is somewhat surprising in that previous work has focused entirely on nilpotent Lie groups.

In §11 we discuss some questions of local geometry. Here the contrasts with Riemannian geometry become very apparent—for example small triangles are not approximately Euclidean, and cut points are always near at hand. Although the results obtained are quite superficial, it appears that there are many interesting results here waiting to be discovered. Further results are given in [39].

The last section is devoted to applications to sub-Laplacian operators (Hormander type sums of squares). It can be said that sub-Riemannian geometry is to the sub-Laplacian what Riemannian geometry is to the Laplacian. Since sub-Laplacian operators are considered popular and respectable mathematics, these applications (and the applications to control theory stressed by Brockett) should indicate that the subject of sub-Riemannian geometry is of more than purely formalistic interest. Nevertheless, these applications are not the raison d'entre of this paper, and in fact only use a small part of the material developed.

The early sections of the paper are devoted to preliminaries. In §2 the important concept of raised Christoffel symbols is introduced (this idea also appears in Gunther's unpublished thesis [17]). In §3 we discuss the length of curves and in §4 geodesics. These sections contain material that is used throughout the paper, as well as some miscellaneous observations and generalizations of standard Riemannian results (for example, the Gauss Lemma).

Notably absent in this work is any notion of covariant derivative and curvature. It appears that it would be barking up the wrong tree to try to distort the Riemannian definitions to make sense in this context. After all, curvature is a measurement of the higher order deviation of the manifold from the Euclidean model, and here there is no approximate Euclidean behavior. Is there an alternative model? This is an extremely interesting question. Brockett [3] has attempted to sketch an approach to it, and Mitchell [30] has computed a "tangent cone" at a point, which might serve as a model space.

Aside from the above-mentioned conjecture for isometries, our results seem to be fairly complete under the strong bracket generating hypothesis. Since this is a rather restrictive assumption, it would be very desirable to extend the results to the general case.

An obvious generalization of sub-Riemannian metrics would be to suppose that the quadratic form is only nondegenerate on S_x. One might call this *sub-Lorentzian geometry*. A number of our results clearly extend to this set-up.

We have tried to keep this paper self-contained, so as not to send the reader hopping from reference to reference (a general background in Riemannian geometry is required, as for example in the first chapter of [24]). Consequently, we present proofs for results which have appeared in other works, without specific attribution. We have attempted to give as complete a bibliography as possible (our apologies to any authors whose works we have overlooked), and we urge the reader to explore the literature to learn about alternate approaches to the subject. From the point of view of control theory, the subject has been studied by Roger Brockett and some of his students ([3], [4], [17], [41], [42]) with earlier work by Carathéodory [6] and Hermann [19], [20], [21]. The special case of the Heisenberg group has been analysed in detail by Koranyi [25], [26]. After this work was completed, the preprint of Hamenstädt [18] appeared which gives a different approach to the theory of geodesics.

2. Preliminaries

Let M be a connected n-dimensional manifold ($n \geqslant 3$) of class C^∞ (C^k for large enough k would suffice). Fix an integer m, $0 < m < n$, called the *subdimension*. Let T_x and T_x^* denote the tangent and cotangent spaces at a point $x \in M$, and $\langle Y, \xi \rangle$ the pairing between them, $Y \in T_x$, $\xi \in T_x^*$. Let S denote a fixed subbundle of the tangent bundle, S_x the fiber over x, of fiber dimension m. S will be said to be *bracket generating* if vector fields which are sections of S together with all brackets span T_x at each point. If $Y \in S_x$ and $\tilde{Y}$ is any section of S passing through Y at x, let $S_x + [Y, S_x]$ denote the subspace of T_x spanned by S_x and all vector fields $[\tilde{Y}, X]$ restricted to x, where X varies over sections of S. Since $[Z, X](x) \in S_x$ if Z is a section of S vanishing at x, it follows that $S_x + [Y, S_x]$ does not depend on the choice of $\tilde{Y}$. Similarly we define bracket(k, Y) inductively by bracket$(2, Y) = S_x + [Y, S_x]$ and bracket$(k, Y) = S_x + [\text{bracket}(k-1, Y), S_x]$ and the definiiton does not depend on the choice of the extension $\tilde{Y}$, and bracket$(2, S_x) = S_x + [S_x, S_x]$, bracket$(k, S_x) = S_x + [\text{bracket}(k-1, S_x), S_x]$. We will say $Y \in S_s$ is a *k-step bracket generator* if bracket$(k, Y) = T_x$. Similarly, S will be said to be *k-step bracket generating* if bracket$(k, S_x) = T_x$ for every x. It is easy to give examples of k-step bracket generating subbundles with no k-step bracket generators. If every nonzero tangent vector in S_x for every $x \in M$ is a 2-step bracket generator we will say S satisfies the *strong bracket generating hypothesis*. Most of the theorems in this paper will be proved under this hypothesis.

A *sub-Riemannian metric* on S (always assumed bracket generating) is a smoothly varying in x positive definite quadratic form Q_x on S_x. Given Q_x, we may define a linear mapping $g(x)$: $T_x^* \to T_x$ as follows: for $\xi \in T_x^*$, the linear mapping $Y \to \langle Y, \xi \rangle$ for $Y \in S_x$ can be represented uniquely as $Y \to Q_x(Y, X)$ for some $X \in S_x$; this X is $g(x)\xi$. More concisely, $g(x)$ and Q_x are related by the identity

$$Q_x(Y, g(x)\xi) = \langle Y, \xi \rangle \quad \text{for all } Y \in S_x. \tag{2.1}$$

Notice that the image of $g(x)$ is exactly S_x. From the properties of Q_x it follows easily that $g(x)$ varies smoothly in x and is symmetric and nonnegative definite, but it is not positive definite since it is not onto. Let N_x denote the null-space of $g(x)$, and $N \subseteq T^*$ the bundle with fibers N_x. Clearly N_x is the annihilator of S_x in T_x^*.

Conversely, given a symmetric nonnegative definite linear operator $g(x)$: $T_x^* \to T_x$ with image S_x, there is a unique positive definite quadratic form Q_x satisfying (2.1). From now on we will assume that the sub-Riemannian metric is given via $g(x)$. In local coordinates we write $g^{jk}(x)$ for the symmetric matrix defining $g(x)$. Note that this is the exact analogue of the raised index metric in Riemannian geometry. There is no analogue of the lowered index metric since $g^{jk}(x)$ is never invertible. As a general rule, any formula of Riemannian geometry that can be expressed in terms of raised indices alone will remain valid in sub-Riemannian geometry. We will use the summation convention of Riemannian geometry.

Lemma 2.1. (a) *If v is a section of the null bundle N, then*

$$g^{jk}\frac{\partial v_k}{\partial x^p} = -\frac{\partial g^{jk}}{\partial x^p}v_k.$$

(b) *If $x(t)$ is a curve in M and $v(t) \in N_{x(t)}$ is a section of N over $x(t)$, then*

$$g^{jk}(x)\dot{v}_k = -\frac{\partial g^{jk}(x)}{\partial x^p}\dot{x}^p v_k$$

for all t (here the dot denotes the t derivative and we have surpressed the t variable throughout).

(c) *If v and w are sections of N, then*

$$\frac{\partial g^{jk}}{\partial x^p}v_j w_k = 0.$$

Proof. To prove (a) apply $\partial/\partial x^p$ to the identity $g^{jk}(x)v_k(x) = 0$ which defines the null bundle. To prove (b) apply d/dt to the identity $g^{jk}(x(t))v_k(t) = 0$. Finally to prove (c) first apply (a) to obtain

$$\frac{\partial g^{jk}}{\partial x^p}v_j w_k = -g^{jk}\frac{\partial v_j}{\partial x^p}w_k$$

(by the symmetry of g^{jk}) and then $g^{jk}w_k = 0$. q.e.d.

We will use these identities throughout without further explanation.

Next we translate the bracket generating property of S into properties of $g(x)$. If $X, Y \in S_x$ and $\tilde{X}$, $\tilde{Y}$ are extensions to sections of S, then $[\tilde{X}, \tilde{Y}]$ at x depends mod S_x only on X and Y, so we may consider $[X, Y]$ as an element of T_x/S_x. Since N_x is the annihilator of S_x in T_x^*, all the pertinent information concerning $[X, Y]$ is contained in the values of $\langle [\tilde{X}, \tilde{Y}], v\rangle$ at x for v varying over N_x. This expression is independent of the extensions and so we will write it as $\langle [X, Y], v\rangle$. Now $X \in S_x$ means there exists $\xi \in T_x^*$ with $X = g(x)\xi$, and similarly $Y = g(x)\eta$. Of course ξ and η are not unique and should be regarded as elements of T_x^*/N_x. Thus we are led to consider the trilinear form $\langle [g\xi, g\eta], v\rangle$ on $(T_x^*/N_x) \times (T_x^*/N_x) \times N_x$.

Lemma 2.2. *In local coordinates*

$$\langle [g\xi, g\eta], v\rangle = \left(g^{jp}\frac{\partial g^{qr}}{\partial x^j} - g^{jq}\frac{\partial g^{rp}}{\partial x^j}\right)\xi_p\eta_q v_r.$$

Proof. Let $\xi(x)$ and $\eta(x)$ denote any sections of T^* extending ξ and η. Then $\tilde{X}^r = g^{rp}(x)\xi_p(x)$ and $\tilde{Y}^r = g^{rq}(x)\eta_q(x)$ are sections of S extending $g\xi$ and $g\eta$, and

$$\begin{aligned}[\tilde{X}, \tilde{Y}]^r &= \tilde{X}^j\frac{\partial}{\partial x^j}\tilde{Y}^r - \tilde{Y}^j\frac{\partial}{\partial x^j}\tilde{X}^r \\ &= g^{jp}\xi_p\frac{\partial g^{qr}}{\partial x^j}\eta_q - g^{jq}\eta_q\frac{\partial g^{rp}}{\partial x^j}\xi_p + g^{jp}\xi_p g^{qr}\frac{\partial \eta_q}{\partial x^j} - g^{jp}\eta_q g^{rp}\frac{\partial \xi_p}{\partial x^j}.\end{aligned}$$

However, on taking the inner product with v, the last two terms are annihilated (because of the factors g^{qr} and g^{rp}) so we obtain the given expression. Note also by the previous lemma this expression is unchanged if ξ or η are changed by a null cotangent. q.e.d.

In order to interpret the lemma we introduce the *raised Christoffel symbols* (also used in [17])

$$\Gamma^{kpq} = \frac{1}{2}\left(g^{jp}\frac{\partial g^{kq}}{\partial x^j} + g^{jq}\frac{\partial g^{kp}}{\partial x^j} - g^{jk}\frac{\partial g^{pq}}{\partial x^j}\right). \tag{2.2}$$

These are the exact analogues of raising indices in the Christoffel symbols Γ_{ij}^k of Riemannian geometry. For $\xi \in T_x^*$ and $v \in N_x$ define $\Gamma(\xi, v) \in T_x$ by $\Gamma^k(\xi, v) = \Gamma^{kpq}\xi_p v_p$. In contrast to the situation in Riemannian geometry where the Christoffel symbols are never tensorial, we have

Lemma 2.3. *$\Gamma(\xi, v)$ is a well-defined tangent vector (independent of the choice of coordinates). In fact $\Gamma(\xi, v) \in S_x$ and $\Gamma(\xi + w, v) = \Gamma(\xi, v)$ for $w \in N_x$, so that Γ: $(T_x^*/N_x) \times N_x \to S_x$.*

Proof. Using Lemma 2.1 we obtain $\Gamma(\xi, v) \in S_x$ and $\Gamma(\xi + w, v) = \Gamma(\xi, v)$ (for example

$$\Gamma^k(\xi, v) = g^{kj}\left(-\frac{1}{2} g^{qk}\xi_p \frac{\partial v_q}{\partial x^j} - \frac{1}{2}\frac{\partial g^{pq}}{\partial x^j}\xi_p v_q \right)\Bigg).$$

To prove that $\Gamma^k(\xi, v)$ transforms as a tangent vector we consider a local diffeomorphism ψ with $\psi(x) = y$ determining a new coordinate system. Then denoting by $\tilde{g}$, $\tilde{\xi}$, and $\tilde{v}$ the expressions for g, ξ, and v in the new coordinates, we have

$$\xi_k = \frac{\partial \psi^j(x)}{\partial x^k}\tilde{\xi}_j, \quad v_k = \frac{\partial \psi^j(x)}{\partial x^k}\tilde{v}_j, \quad \tilde{g}^{kj}(y) = g^{pq}(x)\frac{\partial \psi^j(x)}{\partial x^p}\frac{\partial \psi^j(x)}{\partial x^q}.$$

To see how Γ^{kpq} transforms look at the first term $g^{jp}\partial g^{qk}/\partial x^j$. In the new coordinates

$$\tilde{g}^{jp}\frac{\partial \tilde{g}^{qk}}{\partial x^j} = \left(g^{ab}\frac{\partial y^j}{\partial x^a}\frac{\partial y^p}{\partial x^b} \right)\frac{\partial x^l}{\partial y^j}\frac{\partial}{\partial x^l}\left(g^{cd}\frac{\partial y^q}{\partial x^c}\frac{\partial y^k}{\partial x^d} \right)$$

hence

$$\tilde{g}^{jp}\frac{\partial \tilde{g}^{qk}}{\partial y^j}\tilde{\xi}_p\tilde{v}_q = \left[\frac{\partial y^k}{\partial x^d}\left(g^{ab}\frac{\partial y^j}{\partial x^a}\frac{\partial y^p}{\partial x^b}\frac{\partial x^l}{\partial y^j}\frac{\partial g^{cd}}{\partial x^l}\frac{\partial y^q}{\partial x^c}\frac{\partial y^k}{\partial x^d} \right) \right.$$
$$\left. + g^{ab}\frac{\partial y^j}{\partial x^a}\frac{\partial y^p}{\partial x^b}\frac{\partial x^l}{\partial y^j}g^{cd}\left(\frac{\partial^2 y^q}{\partial x^l \partial x^c}\frac{\partial y^k}{\partial x^d} + \frac{\partial y^q}{\partial x^c}\frac{\partial^2 y^k}{\partial x^l \partial x^d} \right) \right]\tilde{\xi}_p\tilde{v}_p.$$

Notice that the first term is exactly $(\partial y^k/\partial x^d)(g^{jp}(\partial g^{qd}/\partial x^j)\xi_p\eta_q)$, the tangent bundle transformation of $g^{jp}(\partial g^{qk}/\partial x^j)\xi_p\eta_q$, and the last term vanishes since

$$g^{cd}\frac{\partial y^q}{\partial x^c}\tilde{v}_q = g^{cd}v_c = 0.$$

Finally the middle term is symmetric in p and k.

When we examine the transformation of the third term in Γ^{kpq} we find exactly the same expression with p and k interchanged and a minus sign, hence the unwanted terms cancel. Since the middle term in $\Gamma^{kpq}\xi_p v_q$ is zero, we have

$$\tilde{\Gamma}^{kpq}\tilde{\xi}_p\tilde{v}_q = \frac{\partial y^k}{\partial x^d}\Gamma^{dpq}\xi_p v_q,$$

the desired transformation law.

Theorem 2.4. *A tangent vector $X \in S_x$ is a 2-step bracket generator if and only if $\Gamma(\xi, \cdot)\colon N_x \to S_x$ is injective, where $X = g\xi$. In particular, S satisfies the strong bracket generating hypothesis if and only if $\Gamma(\xi, \cdot)\colon N_x \to S_x$ is injective for every nonnull cotangent ξ and every x.*

Proof. By Lemma 2.2 we have

$$\langle [g\xi, g\eta], v\rangle = 2\Gamma(\xi, v)\eta.$$

For X to be a 2-step bracket generator we must obtain all of T_x/S_x in the form $[X, Y] \bmod S_x$ as Y varies over S_x. But N_x is canonically isomorphic to the dual of T_x/S_x, so the surjectivity of $\eta \to [g\xi, g\eta]$ is equivalent to the injectivity of $v \to \Gamma(\xi, v)$.

Remark. If there exists a 2-step generator in S_x, then the 2-step generators in S_x form an open dense subset. To see this observe that the condition of the theorem involves the injectivity of the mapping $\Gamma(\xi, \cdot)$ which depends linearly on ξ. But the injectivity of a matrix can be expressed as the nonvanishing of sums of squares of determinants of minors, hence the condition on ξ becomes $P(\xi) \neq 0$ for a particular polynomial P, and the complement of an algebraic variety $P(\xi) = 0$ is either empty or open and dense.

There does not appear to be any natural notion of covariant derivative on a sub-Riemannian manifold. The closest we can come to it is an analogue of the symmetrized covariant derivative.

Definition 2.5. The *symmetrized covariant derivative* ∇_{sym} of a tangent field Y is defined by

$$\left(\nabla_{\text{sym}} Y\right)^{jk} = g^{jp}\frac{\partial Y^k}{\partial x^p} + g^{kp}\frac{\partial Y^j}{\partial x^p} - Y^p\frac{\partial g^{jk}}{\partial x^p}.$$

Lemma 2.6. ∇_{sym} *is a well-defined differential operator from tensors of rank* $(1,0)$ *to symmetric tensors of rank* $(2,0)$. *Furthermore if* Y *is a section of* S, *say* $Y = g\xi$, *then* $(\nabla_{\text{sym}} Y)^{jk} v_k = 2\Gamma^j(\xi, v)$ *for any null cotangent field* v.

Proof. To see that $\nabla_{\text{sym}} Y$ transforms as a tensor field of rank $(2,0)$ is a straightforward computation that is the same as in the Riemannian case, and it is obviously symmetric in j and k. Also

$$\left(\nabla_{\text{sym}} g\xi\right)^{jk} v_k = g^{jp}\xi_q\frac{\partial g^{kq}}{\partial x^p}v_k - g^{pq}\xi_q\frac{\partial g^{jk}}{\partial x^p}v_k$$

since all the other terms are zero because v is null, and this is exactly $2\Gamma^j(\xi, v)$. q.e.d.

The raised Christoffel symbol also allows us to define a canonical nonnegative quadratic form on the fibers N_x. Namely, for $v \in N_x$, we set $\|v\|_\Gamma$ equal to the Hilbert-Schmidt norm of the operator $\xi \to \Gamma(\xi, v)$ from T_x^*/N_x to S_x (where $g(x)$: $T_x^*/N_x \to S_x$ establishes duality of these vector spaces). If $u_1, \cdots, u_m$ is any orthonormal set in T_x^*, then $\|v\|_\Gamma^2 = \sum_{j=1}^m \sum_{k=1}^m |\langle \Gamma(u_j, v), u_k\rangle|^2$. By Lemma 2.3 this is defined independent of the choice of coordinates. If S is a 2-step bracket generator, then by Lemma 2.2

this quadratic form is positive definite on N_x. In this case we can also define a canonical measure on the manifold. We choose a complement $N_x^\perp$ to N_x in T_x^* to represent T_x^*/N_x and define an inner product $G^{jk}(x)$ on T_x^* by restricting $g^{jk}(x)$ to $N_x^\perp$, taking $\|v\|_\Gamma^2$ on N_x, and making them orthogonal. Of course there is no canonical choice of $N_x^\perp$, but the point is that $\det G^{jk}(x)$ is independent of the choice. Thus the measure $(\det G^{jk}(x))^{-1/2}\,dx$ is canonical. This measure is also described by Brockett [3]. It is not clear, however, whether or not it is significant.

We conclude this section with a brief discussion of the relationship between sub-Riemannian metrics and contact structures. By definition, a *contact structure* on an odd dimensional manifold is a one-form α such that $\alpha \wedge d\alpha \wedge \cdots \wedge d\alpha$ ($(n-1)/2$ factors of $d\alpha$) never vanishes. If we set $N_x = \operatorname{span}\alpha(x)$ and $S_x = N_x^\perp$, then S is a subbundle of TM of codimension one, and we claim S is bracket generating, and in fact the strong bracket generating hypothesis is satisfied. Indeed let X and Y be sections of S, so $\langle X, \alpha\rangle = 0$ and $\langle Y, \alpha\rangle = 0$. Then a simple computation shows $\langle [XY], \alpha\rangle = \langle d\alpha, Y \otimes X\rangle$. To prove the strong bracket generating hypothesis we need to show that for every nonzero X there exists Y such that $\langle [XY], \alpha\rangle \neq 0$ at each point. But if not then $\langle d\alpha, Y \otimes X\rangle = 0$ at a point for all $Y \in S$, and from this it follows that $\alpha \wedge d\alpha \wedge \cdots \wedge d\alpha = 0$ at that point.

Further concepts related to such structures are *C-R structures*, *Levi metrics*, and *chains*. These are discussed in [8], [11], [23], [43].

3. Lengths of curves

Let $x(t)$ be a piecewise C^1 curve in M for $t \in I$, where I is an interval in $\mathbf{R}$. We say $x(t)$ is a *lengthy curve* if $\dot{x} \in S_x$ for every t, where $\dot{x}$ is defined. If $\xi(t) \in T^*_{x(t)}$ is such that $g(x)\xi = \dot{x}$ for every t (where defined) we say $(x(t), \xi(t))$ is a *cotangent lift* of $x(t)$. Clearly, piecewise continuous cotangent lifts exist and are unique modulo sections of the null bundle N_x over the curve. By abuse of notation we will frequently refer to $\xi(t)$ alone as the cotangent lift, suppressing mention of $x(t)$. The *length* of the curve, $L(x)$, is defined by

$$L(x) = \int_I \langle g(x(t))\xi(t), \xi(t)\rangle^{1/2}\,dt$$

and the *energy* by

$$E(x) = \frac{1}{2}\int_I \langle g(x(t))\xi(t), \xi(t)\rangle\,dt.$$

Clearly, these do not depend on the choice of cotangent lift, and agree with the definitions in Riemannian geometry.

It will be necessary, for technical reasons, to also work in the category of locally Lipschitz curves; this is a convenient category because it has a fundamental theorem of calculus, and many curves of interest will automatically belong to it. A continuous curve $x\colon I \to M$ will be called *locally Lipschitz* if for every compact $J \subset I$ and every local coordinate system intersecting $x(J)$ there exists a constant K such that $|x(t_1) - x(t_2)| \leqslant K|t_1 - t_2|$ (where defined) for all t_1, t_2 in J and the distance $|x(t_1) - x(t_2)|$ is measured in the local coordinates. If this condition holds in a set of coordinate systems covering $x(J)$, then it holds for any coordinate system (the constant K may change). A locally Lipschitz curve has a derivative x existing almost everywhere and in the distribution sense as an element of $L^\infty_{\text{loc}}(I)$, and it can be integrated to recover $x(t)$. We say x is a *lengthy locally Lipschitz* curve (abbreviated L^3-curve) if in addition $\dot{x} \in S_x$ for almost every t. The definition of cotangent lift, length, and energy also makes sense for L^3-curves.

We define the distance function $d(P, Q)$ for points P, Q of M to be the infimum of the lengths of all lengthy (piecewise C^1) curves joining P and Q. By Chow's Theorem ([9] or [22]) there always exist such curves, so the distance is finite. It is convenient to compare this distance with the distance function for a Riemannian metric. We will say that a Riemannian metric G on M is a *contraction* of the sub-Riemannian metric Q (or Q is an *expansion* of G) if G restricted to $S \times S$ equals Q. Clearly such contractions always exist—it suffices to find a complementary bundle to S in T, put a positive definite quadratic form on it, and make it orthogonal to S. It is also possible to obtain the sub-Riemannian metric as a limit of a sequence of contractions $G_{(n)}$ (so $g^{jk} = \lim_{n\to\infty} G^{jk}_{(n)}$), and this is a method of obtaining some information about the sub-Riemannian metric ([25], [26]). However, a great deal of information is lost in the limit, so we have favored other techniques.

It is clear from the definition that a lengthy curve has the same length in the Riemannian geometry of a contraction as in the sub-Riemannian geometry, hence

$$d_R(P, Q) \leqslant d(P, Q), \tag{3.1}$$

where d_R denotes Riemannian distance, since the infimum is taken over a larger set of curves. This explains why we use the term "contraction," and also shows $d(P, Q) > 0$ if $P \neq Q$. This implies that d satisfies the axioms for a metric—the other axioms being immediate consequences of the definition. The topology defined by the metrics d and d_R is the same—this follows from any proof of Chow's theorem (e.g. [22, p. 249]), but they will not be equivalent metrics. In §11 we will give a more precise description of the metric d in the case that S is a two-step bracket generator.

Given a metric space (M, d), there is a natural notion of *arc length*. If x: $[a, b] \to M$ is continuous, then

$$L_A(x) = \sup \sum_{j=1}^{N} d\big(x(t_j), x(t_{j-1})\big),$$

where the supremum is taken over all partitions $a = x_0 < x_1 < \cdots < x_N = b$ of the interval. The next lemma shows the consistency of arc length and the previously defined length of a lengthy curve.

Lemma 3.1. *Let $x(t)$ be a piecewise C^1 lengthy curve on a compact interval. Then $L(x) = L_A(x)$.*

Proof. From (3.1) we obtain immediately $L_{AR}(x) \leqslant L_A(x)$, where L_{AR} denotes arc length in the Riemannian metric. However, it is well known that $L_{AR}(x) = L_R(x)$ in Riemannian geometry, and we have already observed $L_R(x) = L(x)$ for lengthy curves, so we have established $L(x) \leqslant L_A(x)$. However, the reverse inequality is easy. If $a = t_0 < t_1 < \cdots < t_N = b$ is any partition of the interval, then the piece of the curve restricted to $[t_{j-1}, t_j]$ is a lengthy curve joining $x(t_{j-1})$ and $x(t_j)$, hence

$$d\big(x(t_{j-1}), x(t_j)\big) \leqslant \int_{t_{j-1}}^{t_j} \langle g(x(t))\xi(t), \xi(t)\rangle^{1/2}\, dt \tag{3.2}$$

by the definition of d. By summing we obtain $\sum_{j=1}^{N} d(x(t_{j-1}), x(t_j)) \leqslant L(x)$, hence $L_A(x) \leqslant L(x)$ when we take the supremum. q.e.d.

A continuous curve x joining P and Q will be called *length minimizing* if $L_A(x) = d(P, Q)$. Clearly any piece of a length minimizing curve is again length minimizing. As in Riemannian geometry, global existence of length minimizing curves will depend on some completeness assumptions, but local existence is guaranteed.

Lemma 3.2. *For every point P there exists $\varepsilon > 0$ such that if $d(P, Q) \leqslant \varepsilon$, then there exists a length minimizing curve x joining P and Q. We may take x to be parametrized by arc length, and then x is a lengthy Lipschitz curve with $L(x) = L_A(x) = d(P, Q)$.*

Proof. Choose ε so that the closed ball B of radius 2ε about P in the Riemannian metric is compact. Now if Q is any point satisfying $d(P, Q) \leqslant \varepsilon$, we can find a sequence of lengthy curves $x_{(k)}$ joining P and Q such that $L_k = L(x_{(k)}) \to d(P, Q)$. Without loss of generality we may assume that each $x_{(k)}$ is parametrized by arc length, so $x_{(k)}(0) = P$, $x_{(k)}(L_k) = Q$, and $L(x_{(k)}|_{[0,t]}) = t$. We may also assume $L_1 \leqslant 2\varepsilon$ and the lengths L_k are decreasing. Since $L(x_{(k)}) = L_R(x_{(k)})$ it follows that all the curves lie in the compact ball B.

We want to apply the Arzela-Ascoli theorem to the family $x_{(k)}$ regarded as mappings of $[0, d(P,Q)]$ into B with the Riemannian metric. We have to establish uniform equicontinuity, but this is easy since

$$(3.3) \quad d_R\big(x_{(k)}(s), x_{(k)}(t)\big) \leqslant L_R\big(x_{(k)}|_{[s,t]}\big) = L\big(x_{(k)}|_{[s,t]}\big) = t - s$$

for $0 \leqslant s < t \leqslant d(P,Q)$.

Let $x(t)$ be a uniform limit (in d_R metric) of a subsequence of $x_{(k)}$. Clearly $x(t)$ is a curve joining P and Q. To see that it is length minimizing it suffices to show $L_A(x) \leqslant d(P,Q)$, since the reverse inequality is automatic. But we can pass to the limit in (3.3) to obtain $d(x(s), x(t)) \leqslant t - s$. Notice this shows $d(x(s), x(t)) = t - s$ so the length minimizing curve is already parametrized by arc-length, and is Lipschitz (in both d and d_R metrics).

Now $\dot{x}(t)$ exists almost everywhere, and it remains to show $\dot{x}(t) \in S_x(t)$ almost everywhere. To do this we examine a difference quotient $h^{-1}(x(t+h) - x(t))$. This is the limit of the difference quotients for the functions $x_{(k)}$. Now we may arrange to take the $x_{(k)}$ to be actually C^1 lengthy curves. We merely lift $x_{(k)}(t)$ to $\xi_{(k)}(t)$, extend $\xi_{(x)}$ to a piecewise continuous section of T^*, approximate by a continuous section $\eta_{(k)}$, and solve the initial value problem $\dot{y}_{(k)} = g\eta_{(k)}$, $y_{(k)}(0) = P$. In the process we lose control of the endpoint of $y_{(k)}$, but we still have $y_{(k)} \to x$ uniformly as $k \to \infty$. Now, by the mean value theorem,

$$h^{-1}\big(y_{(k)}(t+h) - y_{(k)}(t)\big) = \dot{y}_{(k)}(t_k) \in S_{y_{(k)}(t_k)}$$

and so $h^{-1}(x(t+h) - x(t)) \in S_{x(s)}$ for some point in s in $[t, t+h]$. For a point t where $\dot{x}$ exists it follows that $\dot{x}(t) \in S_{x(t)}$.

We can also choose cotangent lifts $\xi_{(k)}(t)$ for $x_{(k)}(t)$ which are uniformly bounded since

$$(3.4) \qquad \big\langle g\big(x_{(k)}(t)\big)\xi_{(k)}(t), \xi_{(k)}(t)\big\rangle = 1$$

follows from (3.3). By passing to a subsequence we can then have $\xi_{(k)}(t)$ converge to $\xi(t)$ in the weak topology of L^∞ dual to L^1. Now $g(x_{(k)}(t))$ converges uniformly to $g(x(t))$ so from

$$x_{(k)}(t) = \int_0^t g\big(x_{(k)}(s)\big)\xi_{(k)}(s)\, ds$$

we can pass to the limit to obtain

$$x(t) = \int_0^t g(x(s))\xi(s)\, ds.$$

This shows $\dot{x}(t) = g(x(t))\xi(t)$ almost everywhere, so $\dot{x}(t) \in S_x(t)$. We can also pass to the limit in (3.4) to obtain $\big\langle g(x(t))\xi(t), \xi(x)\big\rangle = 1$ almost everywhere, hence $L(x) = L_A(x)$.

Corollary 3.3. *Any two points may be joined by a finite sequence of length minimizing curves.*

Proof. Fix a point P and consider the set of all points which can be joined to P by a finite sequence of length minimizing curves. By the lemma it is open (if you can reach Q, apply the lemma to Q) and closed (if Q is a limit point of the set, apply the lemma to connect Q to a point of the set). Since M is connected, the result follows.

4. Geodesics

Given the sub-Riemannian metric $g(x)$: $T_x^* \to T_x$ we can form the Hamiltonian function

$$H(x, \xi) = \tfrac{1}{2}\langle g(x)\xi, \xi\rangle$$

on T^* and consider the Hamilton-Jacobi equation $\dot{x} = \nabla_\xi H$, $\dot{\xi} = -\nabla_x H$ for curves in T^*. Explicitly these equations (abbreviated H-J) are

$$\text{(H-J)} \qquad \begin{aligned} \dot{x}^k(t) &= g^{kj}(x(t))\xi_j(t), \\ \dot{\xi}_k(t) &= -\frac{1}{2}\frac{\partial g^{pq}}{\partial x^k}(x(t))\xi_p(t)\xi_q(t). \end{aligned}$$

A C^2 curve $x(t)$ in M will be called a geodesic if there exists a cotangent lift satisfying (H-J). By abuse of notation we will sometimes refer to the solution of (H-J) as a geodesic. Notice that the first of the (H-J) equations simply says that x is a lengthy curve. As is always the case, (H-J) implies that H is constant along the curve; in this case it means that a geodesic is parametrized by a multiple of arc-length.

Now it is true in Riemannian geometry that geodesics lift to solutions of (H-J) on the cotangent bundle, so we have the correct generalization. Also, if we formulate the variational problem of minimizing energy $E(x)$ over all lengthy curves joining points P and Q over the interval $[0, d(P, Q)]$, then the associated Euler equation is just (H-J). Notice also that if we differentiate the first equation and substitute the second we obtain

$$\ddot{x}^k(t) + \Gamma^k(\xi, \xi) = 0 \tag{4.1}$$

which is the analogue of the geodesic equation in Riemannian geometry. Note, however, that in this case we cannot solve for ξ in terms of x in any obvious way, so (4.1) does not reduce to an equation in x alone; neither is (4.1) together with $\dot{x} = g\xi$ equivalent to (H-J).

The existence and uniqueness theorem for ordinary differential equations says that (H-J) has a unique solution on an interval about zero subject to the initial conditions $x(0) = P$, $\xi(0) = u$ (we will usually assume that coordinates

are chosen so that P is the origin). Furthermore, the solution can be extended until the curve $(x(t), \xi(t))$ either approaches a boundary point of T^* or infinity. However, we can show that the solution can actually be continued as long as $x(t)$ remains in M, so blow-up in the ξ-variable never occurs.

Lemma 4.1. *Let $x(t)$ be a geodesic for $0 \leqslant t < a$ and suppose $x(t)$ remains inside a compact subset of M. Then $x(t)$ can be extended beyond $t = a$.*

Proof. Over the compact set, choose a basis $v^{(1)}(x), \cdots, v^{(n-m)}(x)$ of sections of the null bundle N_x, and complete to a basis of T_x^* by adjoining $u^{(1)}(x), \cdots, u^{(m)}(x)$, all sections bounded and smoothly varying on the compact set. Then write

$$\xi(t) = \sum_{j=1}^{m} a_j(t) u^{(j)}(x(t)) + \sum_{k=1}^{n-m} b_k(t) v^{(k)}(x(t)) \tag{4.2}$$

and consider (H-J) as a system of equations for $x(t)$, $a_j(t)$, and $b_k(t)$. The first key observation is that the functions $a_j(t)$ are uniformly bounded. This follows from the fact that H is constant on the curve and the computation $H(x, \xi) = \frac{1}{2}\Sigma_{j,k} \langle g u^{(j)}, u^{(k)} \rangle a_j a_k$, since the matrix $\langle g u^{(j)}, u^{(k)} \rangle$ is positive definite hence bounded below on a compact set. The second key observation is that when we substitute (4.2) into the second (H-J) equation, all the quadratic terms in b_k vanish, on the left side because they are multiplied by $g v^{(k)}$, and on the right side because of Lemma 2.1(c). The resulting equations are thus

$$\begin{aligned} &\sum_{j=1}^{m} \dot{a}_j u_k^{(j)} + \sum_{j=1}^{n-m} \dot{b}_j v_k^{(j)} \sum_{l=1}^{m} \sum_{l=1}^{m} a_j \frac{\partial u_k^{(j)}}{\partial x^r} g^{rp} a_l u_p^{(l)} \\ &\qquad + \sum_{j=1}^{n-m} \sum_{l=1}^{m} b_j \frac{\partial v_k^{(j)}}{\partial x^r} g^{rp} a_l u_p^{(l)} \\ &\qquad = \frac{1}{2} \frac{\partial g^{pq}}{\partial x^k} \left(\sum_{j=1}^{m} \sum_{l=1}^{m} a_j u_p^{(j)} a_l u_q^{(l)} + \sum_{j=1}^{m} \sum_{l=1}^{n-m} a_j u_p^{(j)} b_l v_q^{(l)} \right). \end{aligned}$$

Now we dot these equations with $v_k^{(1)}, v_k^{(2)}, \cdots, v_k^{(n-m)}$ and solve for b_j. The result is $(n - m)$ linear first order equations in normal form in the $(n - m)$ functions b_j, and the coefficients depend on x^k, a_j, $u^{(j)}$, $v^{(j)}$, hence are bounded. But for linear differential equations with bounded coefficients we have global existence, so the b_j functions are also bounded. Thus all the functions $x^k(t)$ and $\xi_k(t)$ are uniformly bounded, and the local existence theorem implies the solution extends. q.e.d.

With the aid of this lemma we can define a canonical exponential map (we leave it as an exercise to verify that the (H-J) equations transform appropriately under change of variable, so the geodesics do not depend on the

choice of coordinates). Fix any point P in M. For any $u \in T_P^*$, set $\exp_P(u) = x(1)$, where $x(t)$ is the geodesic with cotangent lift satisfying (H-J) with $x(0) = P$, $\xi(0) = u$, provided of course the geodesics extends to $t = 1$. But if $\langle g(P)u, u\rangle \leqslant \varepsilon^2$, then the length of the geodesic on the interval $[0, t]$ will be at most εt, hence the Riemannian distance from P to $x(t)$ will be at most εt. Thus we need only choose ε small enough that the closed Riemannian ball of radius ε about P is compact, in order that $\exp_P(u)$ exist. Thus the exponential map always exists on a cylindrical neighborhood of the origin in T_P^*.

Now the exponential map is always differentiable, since the solution of the system (H-J) depends differentiably on the initial data. But in contrast to the Riemannian case, the exponential map is not a diffeomorphism at the origin. In fact all the geodesics emanating from P must have tangent vectors in S_p. Also it is easy to see that $\exp_P(v) = P$ for any null cotangent $v \in N_p$. The only hope for good behavior is thus at points u near the origin but not null.

We begin with the analogue of the Gauss lemma.

Lemma 4.2. *Let u denote a nonnull cotangent in T_P^* lying inside the cylindrical neighborhood $\langle g(P)u, u\rangle \leqslant \varepsilon^2$, where $\exp_P$ is defined. Let r denote the radial tangent vector and Y any tangent vector orthogonal to r at the point u in T_P^* with respect to the (degenerate) $g(P)$ quadratic form. In other words, $r = \langle g(P)u, u\rangle^{-1/2}u$ and $\langle g(P)r, Y\rangle = 0$. Then*

$$\langle d\exp_P(u)Y, \xi\rangle = 0, \tag{4.3}$$

where ξ is the cotangent lift of the geodesic $t \to \exp_P(tu)$ at $t = 1$.

Proof. Notice that the orthogonality of $d\exp_P(u)Y$ and $d\exp_P(u)r$ with respect to the sub-Riemannian metric $g(\exp_P(u))$ is equivalent to (4.3) if we assume $d\exp_P(u)Y \in S_{\exp_P(u)}$, and then it does not depend on the choice of cotangent lift ξ. This is the case we are interested in. However, in order to prove the result, we do not want to assume $d\exp_P(u)Y \in S_{\exp_p(u)}$, and in this generality we have to specify the cotangent lift given by the (H-J) equations.

The proof is now identical to the Riemannian proof (e.g. [24, p. 79]). Let $u(s)$ be a curve in T_p^* with $u(0) = u$ and

$$\langle g(P)u(s), u(s)\rangle = \text{constant}, \tag{4.4}$$

and set $x(t, s) = \exp_P(tu(s))$. Then

$$\frac{\partial x}{\partial s}(1,0) = d\exp_P(u)\frac{\partial u}{\partial s}(0) = d\exp_P(u)Y$$

if we set $Y = (\partial u/\partial s)(0)$. Note by (4.4) that Y is orthogonal to u hence r, and conversely we can obtain all tangent vectors orthogonal to r in this fashion. Thus (4.3) is equivalent to $\langle(\partial x/\partial s)(1,0), \xi(1,0)\rangle = 0$, where $\xi(t, s)$ is the

cotangent lift of the geodesic $t \to \exp_P(tu(s))$. We will actually prove $f(t) = \langle (\partial x/\partial s)(t,0), \xi(t,0) \rangle = 0$ for $0 \leqslant t \leqslant 1$ by showing $f(0) = 0$ and $f'(t) = 0$. But since $x(0,s) = P$ we have $(\partial x/\partial s)(0,0) = 0$ hence $f(0) = 0$. Also

$$f'(t) = \dot{\xi}_k(t,0)\frac{\partial x^k}{\partial s}(t,0) + \xi_p(t,0)\frac{\partial^2 x^p}{\partial s \partial t}(t,0)$$

and we can use (H-J) to replace $\dot{\xi}_k$ and $(\partial x^p/\partial t)(t,s)$ (for each fixed s, $x(\cdot,s)$ is a geodesic) to obtain

$$\begin{aligned} f'(t) &= -\frac{1}{2}\frac{\partial g^{pq}(x)(t,0)}{\partial x^k}\xi_p(t,0)\xi_q(t,0)\frac{\partial x^k}{\partial s}(t,0) \\ &\quad + \xi_p(t,0)\frac{\partial}{\partial s}\big(g^{pq}(x(t,0))\xi_q(t,0)\big) \\ &= \frac{1}{2}\frac{\partial}{\partial s}\big(g^{pq}(x(t,0))\xi_p(t,0)\xi_q(t,0)\big) \\ &= \frac{1}{2}\frac{\partial}{\partial s}\langle g(x)\xi,\xi\rangle|_{(t,0)}. \end{aligned}$$

However $\langle g(x)\xi,\xi\rangle = \langle g(P)u,u\rangle$ by the invariance of the Hamilton over the (H-J) flow so $f'(t) \equiv 0$.

Lemma 4.3. *Let x: $[a,b] \to M$ be a lengthy Lipschitz curve, and suppose there exists a Lipschitz curve w: $[a,b] \to T_p^*$ such that $x(t) = \exp_P w(t)$, with $w(a) = \lambda w(b)$. Then*

$$L_A(x) \geqslant \left|\langle g(P)w(b),w(b)\rangle^{1/2} - \langle g(P)w(a),w(a)\rangle^{1/2}\right| \tag{4.5}$$

with equality holding if and only if x is a reparametrization of the geodesic $\exp(tw(b))$.

Proof. Write the curve $w(t)$ in polar coordinates $w(t) = r(t)u(t)$, where $\langle g(P)u(t),u(t)\rangle = 1$ and $r(t) \geqslant 0$. Then

$$\begin{aligned} \dot{x}(t) &= \frac{d}{dt}\exp_P(r(t)u(t)) = d\exp_P(w(t))(\dot{r}(t)u(t) + r(t)\dot{u}(t)) \\ &= \dot{r}(t)g(x(t))\xi(t) + d\exp_P(w(t))Y(t), \end{aligned}$$

where $Y(t) = r(t)\dot{u}(t)$ satisfies the hypotheses of the previous lemma, and $\xi(t)$ is the cotangent lift of the geodesic $s \to \exp_P(su(t))$ at $s = r(t)$. The hypothesis that x is lengthy implies that $d\exp(w(t))Y(t) \in S_x(t)$, say $d\exp_P(w(t))Y(t) = g(x(t))\eta(t)$ a.e. Then the length of x is obtained by integrating

$$\langle \dot{r}(t)\xi(t) + \eta(t),\ \dot{r}(t)g(x(t))\xi(t) + g(x(t))\eta(t)\rangle^{1/2}$$

which equals

$$\Big(|\dot{r}(t)|^2 + \langle g(x(t))\eta(t),\eta(t)\rangle\Big)^{1/2} \tag{4.6}$$

since $\langle g(x(t))\xi(t), \xi(t)\rangle = 1$ and the cross terms are zero by the previous lemma. Clearly (4.6) dominates $|\dot{r}(t)|$ which gives (4.5), and equality holds if and only if $r(t)$ is monotone and $d\exp_P(w(t))Y(t) = 0$. But then x satisfies the differential equation $\dot{x}(t) = \dot{r}(t)g(x(t))\xi(x(t))$ and so does $\exp_P(r(t)u(t_0))$ for any fixed t_0. Thus $u(t)$ is constant and x is a reparametrized geodesic. q.e.d.

We have observed that a general lengthy curve does not have a unique cotangent lift. If the curve is a geodesic, however, there is a special cotangent lift, the one that solves (H-J). It is natural to ask if there is in general a canonical cotangent lift which makes the correct choice for geodesics. The following result gives an answer under the strong bracket generating hypothesis.

Lemma 4.4. *Assume the strong bracket generating hypothesis. Let $x(t)$ be any Lipschitz lengthy curve. Then there is a unique cotangent lift $(x(t), \xi(t))$ with the property that*

$$\dot{\xi}_j + \frac{1}{2}\frac{\partial g^{pq}(x)}{\partial x^j}\xi_p\xi_q$$

is orthogonal to all $\Gamma(\xi, v(x))$ for every $v \in N_x$ at a.e. t. The lift is independent of the choice of coordinates, and is called the canonical lift.

Proof. Start with any cotangent lift $(x(t), \eta(t))$, and let $v^{(1)}, \cdots, v^{(n-m)}$ be a basis for sections of N over a neighborhood of the curve. Then the most general cotangent lift is of the form

$$\xi(t) = \eta(t) + \sum_{k=1}^{n-m} a_k(t)v^{(k)}(x(t)) = \eta(t) + v(t).$$

Now

$$\dot{\xi}_j + \frac{1}{2}\frac{\partial g^{pq}(x)}{\partial x^j}\xi_p\xi_q = \eta_j + \frac{1}{2}\frac{\partial g^{pq}(x)}{\partial x^j}\eta_p\eta_q + \left(\dot{v}_j + \frac{\partial g^{pq}(x)}{\partial x^j}\eta_p v_q\right).$$

But recall that (for $w \in N_x$)

$$\Gamma^k(\xi, w) = g^{kj}\left(\dot{w}_j + \frac{\partial g^{pq}(x)}{\partial x^j}\xi_p w_q\right)$$

and ξ and η are interchangeable here. Thus the orthogonality condition

$$\left(\xi_j + \frac{1}{2}\frac{\partial g^{pq}(x)}{\partial x^j}\xi_p\xi_q\right)\Gamma^j(\xi, w) = 0$$

amounts to $n - m$ linear equations in the $n - m$ variables $a_k(t)$ at each point t, and the equations are uniquely soluble since $\Gamma(\xi, \cdot)$: $N \to S$ is injective. Finally we have already obesrved that $\Gamma^k(\xi, w)$ transforms as a tangent vector,

and it is a straightforward exercise to verify that

$$\dot{\xi}_j + \frac{1}{2}\frac{\partial g^{pq}(x)}{\partial x^j}\xi_p\xi_q$$

transforms as a cotangent vector under change of variable.

5. Derivative of exp

We want to compute $d\exp_P$ in order to show that the exponential mapping is a local diffeomorphism at some points. To do this we compute the Taylor expansion of $\exp_P$ about the origin. Let us fix P at the origin of coordinates. Then

$$\exp_P(u)^k = \sum_{r=1}^{N} \frac{1}{r!}\gamma_{(r)}^{kp_1\cdots p_r}u_{p_1}u_{p_2}\cdots u_{p_r} + O\big(|u|^{N+1}\big), \tag{5.1}$$

where the $\gamma_{(r)}^{kp_1\cdots p_r}$ are symmetric in $(p_1\cdots p_r)$ and can be computed in terms of g and its derivatives at the origin. Now $\exp_P(tu) = x(t)$, where $(x(t), \xi(t))$ is the solution of (H-J) with $x(0) = 0$, $\xi(0) = u$, so $\gamma_{(r)}^{kp_1\cdots p_r}u_{p_1}\cdots u_{p_r} = (d/dt)^r x(0)$. Using (H-J) we obtain the recursion relation

$$\begin{aligned}\left(\frac{d}{dt}\right)^{r+1}x^k(t) &= \frac{d}{dt}\big(\gamma_{(r)}^{kp_1\cdots p_r}(x(t))\xi_{p_1}(t)\cdots\xi_{p_r}(t)\big)\\ &= \left(\frac{\partial\gamma_{(r)}^{kp_1\cdots p_r}}{\partial x^q}(x(t))g^{qp_{r+1}}(x(t))\right.\\ &\qquad\left. - \frac{r}{2}\gamma_{(r)}^{kp_1\cdots p_{r-1}q}\frac{\partial g^{p_rp_{r+1}}}{\partial x^q}(x(t))\right)\xi_{p_1}(t)\cdots\xi_{p_{r+1}}(t).\end{aligned}$$

Hence we have

$$\begin{aligned}&\gamma_{(r+1)}^{kp_1\cdots p_{r+1}}(x) = \operatorname{sym}(p_1,\cdots,p_{r+1})\\ &\cdot\left(g^{qp_{r+1}}(x)\frac{\partial\gamma_{(r)}^{kp_1\cdots p_r}(x)}{\partial x^q} - \frac{r}{2}\gamma_{(r)}^{kp_1\cdots p_{r-1}q}(x)\frac{\partial g^{p_rp_{r+1}}}{\partial x^q}(x)\right),\end{aligned} \tag{5.2}$$

where $\operatorname{sym}(p_1,\cdots,p_{r+1})$ means we symmetrize in the indices $p_1,\cdots,p_{r+1}$. From (H-J) we get

$$\gamma_{(1)}^{kp} = g^{kp} \tag{5.3}$$

and we have already observed that

$$\gamma_{(2)}^{kp_1p_2} = -\Gamma^{kp_1p_2}. \tag{5.4}$$

The general term is clearly quite complicated, but it will be sufficient for us to understand some aspects of $\gamma_{(3)}$. By differentiating (5.1) we obtain the Taylor expansion

$$(5.5)\quad d\exp_P(u)^{kj} = g^{kj}(0) + \sum_{r=2}^{N} \frac{1}{(r-1)!}\gamma_{(r)}^{kjp_2\cdots p_r}u_{p_2}\cdots u_{p_r} + O\big(|u|^N\big).$$

Now let us assume the strong bracket generating hypothesis. We choose coordinates near P so P is the origin and

$$g^{jk}(0) = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

where I is the $m \times m$ identity matrix. It is convenient to adopt the following convention: lower case Latin letters $a, b, \cdots$ from the beginning of the alphabet denote indices ranging from $1, \cdots, m$, while corresponding lower case Greek letters $\alpha, \beta, \cdots$ denote indices ranging from $m+1, \cdots, n$. Thus a typical $n \times n$ matrix will be written

$$(5.6)\qquad M^{jk} = \begin{pmatrix} A^{ab} & B^{a\beta} \\ C^{\alpha b} & D^{\alpha\beta} \end{pmatrix}.$$

If we write $d\exp_P(u)$ in this form and compute the leading terms of A, B, C, D we find

$$(5.7)\qquad \begin{aligned} A &= I + O(|u|), \\ B^{a\beta} &= -\Gamma^{a\beta p}u_p + O\big(|u|^2\big), \\ C^{\alpha b} &= -\Gamma^{\alpha b p}u_p + O\big(|u|^2\big), \\ D^{\alpha\beta} &= \tfrac{1}{2}\gamma_{(3)}^{\alpha\beta pq}u_p u_q + O\big(|u|^3\big) \end{aligned}$$

as a consequence of (5.3)–(5.5) (recall that $\gamma^{\alpha\beta p} = 0$ so there are no terms of order two in D). Here $|u|$ denotes any Euclidean norm on T_p^*. Now a simple determinant computation shows

Lemma 5.1. *If $M(u)$ is any matrix given by* (5.6) *with* (5.7), *then* $\det M(u) = \det \tilde{M}(u) + O(|u|^{2(n-m)+1})$, *where $\tilde{M}$ is obtained from M by discarding the error terms in* (5.7), *and* $\det \tilde{M}(u)$ *is homogeneous of degree* $2(n-m)$ *in u.*

Now to compute $\det \tilde{M}(u)$ we need to compute $\gamma_{(3)}^{\alpha\beta pq}$. By (5.2) and (5.4)

$$\begin{aligned} \gamma_{(3)}^{\alpha\beta pq} = \frac{1}{3}\Bigg(&\Gamma^{\alpha\beta r}(x)\frac{\partial g^{pq}}{\partial x^r}(x) + \Gamma^{\alpha pr}(x)\frac{\partial g^{\beta q}(x)}{\partial x^r} + \Gamma^{\alpha qr}\frac{\partial g^{p\beta}(x)}{\partial x^r} \\ &- g^{rq}(x)\frac{\partial \Gamma^{\alpha\beta p}(x)}{\partial x^r} - g^{rp}(x)\frac{\partial \Gamma^{\alpha\beta q}(x)}{\partial x^r} - g^{r\beta}(x)\frac{\partial \Gamma^{\alpha pq}(x)}{\partial x^r}\Bigg), \end{aligned}$$

and setting $x = 0$ we obtain

$$(5.8)\qquad \gamma_{(3)}^{\alpha\beta ab} = \frac{1}{3}\Bigg(\Gamma^{\alpha ac}\frac{\partial g^{\beta b}}{\partial x^c} + \Gamma^{\alpha bc}\frac{\partial g^{\beta a}}{\partial x^c} - \frac{\partial \Gamma^{\alpha\beta a}}{\partial x^b} - \frac{\partial \Gamma^{\alpha\beta b}}{\partial x^a}\Bigg)$$

and $\gamma_{(3)}^{\alpha\beta pq} = 0$ if $p > m$ or $q > m$. Since

$$\Gamma^{\alpha\beta a}(x) = \tfrac{1}{2}\left(g^{\alpha j}(x)\frac{\partial g^{\beta a}(x)}{\partial x^j} - g^{\beta j}(x)\frac{\partial g^{\alpha a}(x)}{\partial x^j} - g^{aj}(x)\frac{\partial g^{\alpha\beta}(x)}{\partial x^j}\right)$$

we have at $x = 0$

$$\Gamma^{\alpha\beta a} = -\frac{1}{2}\frac{\partial g^{\alpha\beta}}{\partial x^a} \quad \text{and} \quad \frac{\partial \Gamma^{\alpha\beta a}}{\partial x^b} = \frac{1}{2}\left(\frac{\partial g^{\alpha c}}{\partial x^b}\frac{\partial g^{\beta a}}{\partial x^c} - \frac{\partial g^{\beta c}}{\partial x^b}\frac{\partial b^{\alpha a}}{\partial x^c} - \frac{\partial^2 g^{\alpha\beta}}{\partial x^a \partial x^b}\right)$$

since $\partial g^{\alpha\beta}/\partial x^j = 0$ by Lemma 2.1, and also

$$\Gamma^{abc} = -\frac{1}{2}\left(\frac{\partial g^{ac}}{\partial x^b} + \frac{\partial g^{ab}}{\partial x^c}\right) \quad \text{and} \quad \Gamma^{\alpha\beta c} = \frac{1}{2}\left(\frac{\partial g^{\beta c}}{\partial x^a} - \frac{\partial g^{\alpha\beta}}{\partial x^c}\right).$$

To simplify $\partial^2 g^{\alpha\beta}/\partial x^a \partial x^b$ we use the method of Lemma 2.1. Differentiating part (a) of that lemma we have

$$\frac{\partial^2 g^{jk}(x)}{\partial x^q \partial x^r} v_k(x) = -\frac{\partial g^{jk}(x)}{\partial x^r}\frac{\partial v_k(x)}{\partial x^q} - \frac{\partial g^{jk}(x)}{\partial x^q}\frac{\partial v_k(x)}{\partial x^r} - g^{jk}(x)\frac{\partial^2 v_k(x)}{\partial x^q \partial x^r}$$

for any null-section $v(x)$. Taking the inner product with another null-section $w(x)$ we obtain

$$\frac{\partial^2 g^{jk}}{\partial x^q \partial x^r} v_k(x) w_j(x) = -\frac{\partial g^{jk}(x)}{\partial x^r}\frac{\partial v_k(x)}{\partial x^q} w_j(x) - \frac{\partial g^{jk}(x)}{\partial x^q}\frac{\partial v_k(x)}{\partial x^r} w_j(x).$$

Now set $x = 0$ and take $v_k(x) = \delta_{k\beta}$, $w_j(0) = \delta_{j\alpha}$. Since

$$\frac{\partial v_c}{\partial x^r}(0) = -\frac{\partial g^{kc}}{\partial x^r}(0) v_k$$

by Lemma 2.1(a) and the special form of $g(0)$, we obtain

$$\frac{\partial^2 g^{\alpha\beta}}{\partial x^q \partial x^r} = \frac{\partial g^{\alpha c}}{\partial x^q}\frac{\partial g^{\beta c}}{\partial x^r} + \frac{\partial g^{\alpha c}}{\partial x^r}\frac{\partial g^{\beta c}}{\partial x^q},$$

where we take the summation convention with respect to c, $c = 1, \cdots, m$, despite the fact that it is twice raised. Substituting all these computations back in (5.8) we obtain

$$\begin{aligned}\gamma_{(3)}^{\alpha\beta ab} = \frac{1}{6}\Bigg[&-\left(\frac{\partial g^{\alpha c}}{\partial x^a} + \frac{\partial g^{\alpha a}}{\partial x^c}\right)\frac{\partial g^{\beta b}}{\partial x^c} - \left(\frac{\partial g^{\alpha c}}{\partial x^b} + \frac{\partial g^{\alpha b}}{\partial x^c}\right)\frac{\partial g^{\beta a}}{\partial x^c} \\ &- \frac{\partial g^{\alpha c}}{\partial x^b}\frac{\partial g^{\beta a}}{\partial x^c} + \frac{\partial g^{\beta c}}{\partial x^b}\frac{\partial g^{\alpha a}}{\partial x^c} - \frac{\partial g^{\alpha c}}{\partial x^a}\frac{\partial g^{\beta b}}{\partial x^c} + \frac{\partial g^{\beta c}}{\partial x^a}\frac{\partial g^{\alpha b}}{\partial x^c} \\ &+ 2\frac{\partial x^{\alpha c}}{\partial x^a}\frac{\partial g^{\beta c}}{\partial x^b} + 2\frac{\partial g^{\alpha c}}{\partial x^b}\frac{\partial g^{\beta c}}{\partial x^a}\Bigg].\end{aligned}$$

To simplify this we introduce the abbreviations

$$E^{\alpha\beta} = \frac{\partial g^{\alpha\beta}}{\partial x^b} u_b, \qquad F_a^\beta = \frac{\partial g^{\beta b}}{\partial x^a} u_b$$

and compute

$$\begin{aligned}\gamma_{(3)}^{\alpha\beta ab} u_a u_b &= \tfrac{1}{3}\big(2E^{c\alpha}E^{c\beta} - 2E^{c\alpha}F_c^\beta + E^{c\beta}F_c^\alpha - F_c^\alpha F_c^\beta\big) \\ &= \tfrac{1}{6}\big(F_c^\beta - E^{c\beta}\big)\big(\big(F_c^\alpha - E^{c\alpha}\big) - 3\big(E^{c\alpha} + F_c^\alpha\big)\big) \\ &\quad - \tfrac{2}{3}\tilde{B}^{c\beta}\tilde{B}^{c\alpha} + 2\tilde{B}^{c\beta}\tilde{C}^{\alpha c}.\end{aligned}$$

Thus we have

$$\tilde{M} = \begin{pmatrix} I & \tilde{B}^{a\beta} \\ \tilde{C}^{\alpha b} & \tfrac{1}{3}\tilde{B}^{c\beta}\tilde{B}^{c\alpha} + \tilde{B}^{c\beta}\tilde{C}^{\alpha c} \end{pmatrix}$$

and hence

$$\begin{pmatrix} I & 0 \\ -\tilde{C}^{\alpha b} & I \end{pmatrix} \tilde{M} = \begin{pmatrix} I & \tilde{B}^{a\beta} \\ 0 & \tfrac{1}{3}\tilde{B}^{c\beta}\tilde{B}^{c\alpha} \end{pmatrix}$$

from which we obtain

$$\det \tilde{M} = \det \tfrac{1}{3}\tilde{B}^{c\beta}\tilde{B}^{c\alpha}. \tag{5.9}$$

Lemma 5.2. *Using the strong bracket generating hypothesis, there exists* $\varepsilon > 0$ *such that*

$$\det \tilde{M}(u) \geqslant \varepsilon \langle gu, u\rangle^{(n-m)}. \tag{5.10}$$

Proof. By Theorem 2.4 the mapping $\Gamma(u, \cdot)$: $N \to T$ is injective for every u with $\langle gu, u\rangle \neq 0$. On the other hand, by (5.7), $\tilde{B}^{c\beta}$ is the matrix of $-\Gamma(u, \cdot)$, hence $\tilde{B}^{c\beta}\tilde{B}^{c\alpha}$ is the matrix of $\Gamma(u, \cdot)^{\mathrm{tr}}\, \Gamma(u, \cdot)$ which is invertible since $\Gamma(u, \cdot)$ is injective. Thus $\det \tilde{M}(u) \neq 0$ if $\langle gu, u\rangle \neq 0$ and (5.10) follows by a homogeneity argument.

Remark. The argument shows that $\det \tilde{M}(u) \neq 0$ if and only if gu is a two-step bracket generator.

Theorem 5.3. (a) *If gu is a 2-step bracket generator, then there exists $\varepsilon > 0$ such that $\exp_P(tu)$ is a local diffeomorphism for all t such that $0 < t < \varepsilon$.*

(b) *Under the strong bracket geometry hypothesis, there exists $\varepsilon > 0$ (depending continuously on P) such that $\exp_P(u)$ is a local diffeomorphism provided $gu \neq 0$ and $|u| < \varepsilon(\langle gu, u\rangle/|u|^2)^{(n-m)}$.*

Proof. By Lemma 5.1

$$\det M(u) \geqslant \det \tilde{M}(u) - C|u|^{2(n-m)+1}$$

for u near zero, so part (b) is an immediate consequence of Lemma 5.2, and (a) follows by the remark.

Theorem 5.4. *Let $x(t)$ be any geodesic such that $\dot{x}(0)$ is a 2-step bracket generator. Then it is the unique length minimizing curve joining $x(0)$ and $x(t_0)$ for all sufficiently small t_0. In particular, under the strong bracket generating hypothesis, every nonconstant geodesic is locally a unique length minimizing curve.*

Proof. Choose $P = x(-t_1)$, where t_1 is to be chosen, and write $x(t) = \exp_P((t + t_1)u_1)$. By taking t_1 small enough we can arrange that $gu_1 = \dot{x}(-t_1)$ is a 2-step bracket generator and $\exp_P$ is a local diffeomorphism at t_1u_1. Let U be a neighborhood of t_1u_1 in T_P^* on which $\exp_P$ is a diffeomorphism, let $V = \exp_P(U)$, and let Log: $V \to U$ denote the inverse of $\exp_P$. Choose ε small enough that V contains the Riemannian ball of radius ε about $x(0)$. Then any lengthy curve of length $< \varepsilon$ must remain in V. If $|t_0| < \varepsilon\langle gu, u\rangle^{-1/2}$, then we claim $x(t)$ is the unique length minimizing curve joining $x(0)$ and $x(t_0)$. Indeed let $y(\varepsilon)$ be any lengthy curve joining $x(0)$ and $x(t_0)$ parametrized by arc length $< \varepsilon$. Then $y(t)$ is Lipschitz and $w(t) = \operatorname{Log} y(t)$ is well defined and Lipschitz. Lemma 4.3 then implies that $y(t)$ has length at least as much as $x(t)$, with equality holding if and only if the curves coincide.

Remark. Theorem 5.3(b) is not true in general. If we take the Cartesian product $M_1 \times M_2$ of two sub-Riemannian manifolds, then we can make it into a sub-Riemannian manifold in an obvious way. Of course it will never satisfy the strong bracket generating hypothesis. If we choose $u_1 \times v_2$ in T_P^*, where $g_1u_1 \neq 0$ but $g_2v_2 = 0$, then $g(u_1 \times v_2) \neq 0$ but is easy to see that $\exp_P$ will not be a local diffeomorphism at $u_1 \times v_2$, because varying v_2 in the null directions will cause no change in the geodesic. Of course Theorem 5.4 might still be true in general, but if so it will require a different proof.

6. Length minimizing curves

In this section we show in general that all length minimizing curves are geodesics. This will be a simple consequence of the well-known theorem of Pontryagin giving necessary conditions for the existence of minima to Lagrange problems in the calculus of variations.

First we observe that the problem of minimizing length is essentially equivalent to that of minimizing energy with the domain of the curve fixed (say $[0, 1]$). The reason is the same as in Riemannian geometry: among all the parametrizations of a curve on $[0, 1]$, the one that minimizes energy E is the one with parameter proportional to arc length, i.e. $\langle g(x(t))\xi(t), \xi(t)\rangle = L^2$ a.e., in which case $E = \frac{1}{2}L^2$ [24]. Thus if $x(t)$ is a length minimizing curve on $[0, 1]$ joining P and Q which is parametrized by a multiple of arc length, then

$x(t)$ also minimizes energy among all Lipschitz lengthy curves in $[0,1]$ joining P and Q

To state Pontryagin's theorem we need to define the function

$$H(x,\xi,\lambda) = \tfrac{1}{2}\lambda_0 g^{jk}\xi_j\xi_k + \lambda_j g^{jk}(x)\xi_k,$$

where $\lambda = (\lambda_0, \lambda_1, \cdots, \lambda_n) \in \mathbf{R}^{n+1}$, and then $M(x,\lambda) = \inf_\xi H(x,\xi,\lambda)$. (Here we have chosen a coordinate neighborhood containing the curve, which may require taking only a piece of the curve at a time.) A simple computation completing the square shows

$$(M,\lambda) = \begin{cases} -\lambda_0^{-1} g^{jk}(x)\lambda_j\lambda_k & \text{if } \lambda_0 \neq 0, \\ -\infty & \text{if } \lambda_0 = 0. \end{cases}$$

Theorem 6.1 (*Pontryagin*). *Let $x(t)$ be an energy minimizing curve on $[0,1]$ joining P to Q and $(x(t),\xi(t))$ a cotangent lift. Then there exists a nonvanishing absolutely continuous function $\lambda(t) = (\lambda_0, \lambda_1(t), \cdots, \lambda_n(t))$ with λ_0 constant and $\lambda_0 \geqslant 0$, such that*

(a) $\dot{\lambda}_j = -(\partial H/\partial x^j)(x(t),\xi(t),\lambda(t))$ *a.e.*

(b) $H(x(t),\xi(t),\lambda(t)) = M(x(t),\lambda(t))$ *a.e.*

(c) $M(x(t),\lambda(t))$ *is constant on* $[0,1]$.

This theorem is a special case of [7, Theorem 5.1.i].

Corollary 6.2. *Every length minimizing curve is a geodesic.*

Proof. From (b) we see that $\lambda_0 \neq 0$, and we may take $\lambda_0 = 1$ for simplicity by homogeneity. Then (b) says $g^{jk}(x(t))\xi_k(t) = -g^{jk}(x(t))\lambda_k(t)$, so $(x(t), -\lambda(t))$ is another cotangent lift. Now (a) says

$$-\dot{\lambda}_j \frac{1}{2}\frac{\partial g^{pq}}{\partial x^j}(x)\xi_p\xi_q + \frac{\partial g^{pq}}{\partial x^j}(x)\lambda_p\xi_q$$

and by Lemma 2.1 this is

$$-\dot{\lambda}_j = -\tfrac{1}{2}\frac{\partial g^{pq}}{\partial x^j}(x)(-\lambda_p)(-\lambda_q)$$

so $(x(t), -\lambda(t))$ is a geodesic. Note that (c) says merely that $x(t)$ is parametrized by a multiple of arc length.

Corollary 6.3. *Let $x(t)$ be a length minimizing curve on $[0,1]$ joining P and Q. Then for any $t_0 < 1$, the restriction of $x(t)$ to $[0,t_0]$ is the unique length minimizing geodesic joining P and $x(t_0)$.*

Proof. Suppose $y(t)$ were a distinct length minimizing curve joining P and $x(t_0)$, say parametrized on $[0,t_0]$. Then consider the curve on $[0,1]$

$$z(t) = \begin{cases} y(t) & \text{if } 0 \leqslant t \leqslant t_0, \\ x(t) & \text{if } t_0 \leqslant t \leqslant 1, \end{cases}$$

obtained by following y from P to $x(t_0)$, and then following x from $x(t_0)$ to Q. Since z has the same length as x, it is length minimizing so by Corollary 6.2 it must be a geodesic. But the uniqueness of solutions of (H-J) implies $y = x$ after all. q.e.d.

We can now show that the exponential map must be a local diffeomorphism at some points.

Theorem 6.4. *Let P be any point. Then the set of cotangents u such that $\exp_P$ is a local diffeomorphism in a neighborhood of u, and $\exp_P(tu)$ on $[0,1]$ is a length minimizing geodesic, is a nonempty set of positive measure.*

Proof. Let U be the subset of $u \in T_P^*$ such that $\exp_P(tu)$ on $[0,1]$ is a length minimizing geodesic. Then $\exp_P$ maps U onto the ball of radius one about P, so there must be a subset of U of positive measure on which the determinant of $d\exp_P$ does not vanish.

7. Completeness

We say M is *complete* if it is complete as a metric space. We will prove the analogue of the Hopf-Rinow theorem, relating completeness to extendibility of geodesics. In one direction we need the strong bracket generating hypothesis.

Theorem 7.1. (a) *If M is complete, then every geodesic can be extended indefinitely, and any two points can be joined by a geodesic.*

(b) *Assume the strong bracket generating hypothesis. If there exists a point P such that every geodesic from P can be indefinitely extended, then M is complete. (Recall that we assumed M was connected.)*

Proof. (a) Let $\gamma(t)$ be a geodesic on the interval $0 \leqslant t < T$, parametrized by arc length. Then $d(\gamma(t_1), \gamma(t_2)) \leqslant |t_2 - t_1|$, so by completeness there exists a point P such that $\lim_{t \to T} \gamma(t)) = P$. By so extending γ to $[0, T]$ we obtain a continuous mapping, so the image is compact in M. Then Lemma 4.1 shows how to extend γ past T. Finally we can repeat the proof of Lemma 3.2 to establish the existence of length minimizing curves joining any two points, and these must be geodesics by Corollary 6.2.

(b) We begin by showing that every point of M can be joined to P by a length minimizing geodesic. The proof is essentially the same as in the Riemannian case, and is due to de Rahm [24, p. 126]. We start by using Lemma 3.2 to find $\varepsilon > 0$ so that if $d(P,Q) \leqslant 2\varepsilon$, then the length minimizing geodesic exists. Given any Q in M, we use the compactness of the ε-sphere about P to find R on the sphere such that the distance add, $d(P,Q) = \varepsilon + d(R,Q)$, and we choose a length minimizing geodesic $\gamma(t)$ from P to R parametrized by arc length, and use the hypothesis to extend $\gamma(t)$ up to

$t = d(P, Q)$. We claim this geodesic ends at Q, $\gamma(d(P, Q)) = Q$. To prove this we let T be the supremum of all t such that distances add, $t + d(\gamma(t), Q) = d(P, Q)$. We already know $T \geqslant \varepsilon$, and if $T = d(P, Q)$ this will show $d(\gamma(d(P, Q)), Q) = 0$ as desired. But suppose $T < d(P, Q)$. Let $Q' = \gamma(t)$ so that we have $T + d(Q', Q) = d(P, Q)$. By repeating the original argument for Q' in place of P we can find $\varepsilon' > 0$ and R' such that $d(Q', R') = \varepsilon'$ and $\varepsilon' + d(R', Q) = d(Q', Q)$ and there exists a length minimizing geodesic γ' joining Q' to R'. But the curve from P to R' going first along γ to Q' and then γ' is length minimizing, hence a geodesic. Thus γ' coincides with the continuation of γ, and we have contradicted the maximal property of T.

Now we can prove completeness. Given a Cauchy sequence $\{x_j\}$, we consider a sequence $\{\gamma_j\}$ of length minimizing geodesics parametrized by arc length joining P to x_j. Now we can use an Arzela-Ascoli theorem argument, just as in the proof of Lemma 3.2, to show that, by passing to a subsequence if necessary, there exists a uniform limit $\gamma(t) = \lim_{j \to \infty} \gamma_j(t)$ on a small interval $0 \leqslant t \leqslant \varepsilon$. Now $\gamma(t)$ is length minimizing, hence a geodesic, so $\gamma(t) = \exp_P(tu)$ for some unit cotangent vector u. Similarly $\gamma_j(t) = \exp_P(tu_j)$ for some unit cotangent vector u_j.

We need to show that u is the limit of $\{u_j\}$. This is not obvious because the unit sphere in the cotangent space is not compact, and it is here that we need to use the strong bracket generating hypothesis. Indeed by Theorem 5.3(b), there exists $t_0 > 0$ small enough that $\exp_P$ is a local diffeomorphism in a neighborhood of $t_0 u$. Then by using the uniqueness of the length minimizing geodesics $\gamma_j(t)$ on $0 \leqslant t \leqslant t_0$ (either by Corollary 6.3 or Theorem 5.4) we obtain $t_0 u$ as the limit of $t_0 u_j$. Finally we use the continuity of $\exp_P$ to obtain $\lim_{j \to \infty} \exp_P(t_j u_j) = \exp_P(Tu)$ if $t_j \to T$, and choosing $t_j = d(P, x_j)$ we have $\lim_{j \to \infty} x_j = \exp_P(Tu)$ proving completeness. q.e.d.

As a corollary of the proof we have

Corollary 7.2. *Assume the strong bracket generating hypothesis, and let $T > 0$ be such that the closed ball of radius T about the point P is complete. Then the subset of the unit sphere in the cotangent space at P of all u such that the geodesic $\exp_P(tu)$ on $[0, T]$ is length minimizing, is compact.*

Because the topology of M is locally Euclidean, the completeness of M is equivalent to the compactness of all closed balls. For applications to analysis, the following existence of "approximate constants" is useful:

Theorem 7.3. *The completeness of M is equivalent to the existence of a sequence of functions φ_j: $M \to \mathbf{R}$ satisfying*

(i) *φ_j has compact support,*

(ii) *$\lim_{j \to \infty} \varphi_j(x) = 1$ pointwise for each $x \in M$,*

(iii) *$|\varphi_j(x) - \varphi_j(y)| \leqslant \varepsilon_j d(x, y)$ for all $x, y \in M$ for a sequence $\varepsilon_j \to 0$.*

Proof. Assume M is complete. Take $\varphi_j(x) = h_j(d(P,x))$, where h_j is the real function taking value one on $[0, j]$ and zero on $[2j, \infty]$ and linear in between. Then (i) follows from completeness, (ii) is obvious, and (iii) follows from

$$|\varphi_j(x) - \varphi_j(y)| \leqslant j^{-1}|d(P,x) - d(P,y)| \leqslant j^{-1}d(x,y).$$

Conversely, suppose such functions exist, and let $\{x_k\}$ be a Cauchy sequence. Choose j large enough that $\varphi_j(x_1)$ is close to one, say $\varphi_j(x_1) \geqslant \frac{1}{2}$, and so that $\varepsilon_j d(x_1, x_k) \leqslant \frac{1}{4}$ for all k (since $\{x_k\}$ is Cauchy, $d(x_1, x_k)$ is bounded). Then by (iii) we have $\varphi_j(x_k) \geqslant \frac{1}{4}$ for all k, so the sequence $\{x_k\}$ lies in the compact support of φ_j. Then the compactness shows that $\{x_j\}$ has a limit. q.e.d.

The following result gives a useful criterion for completeness.

Theorem 7.4. *Let M be a sub-Riemannian manifold. If there exists a Riemannian contraction of the metric with respect to which M is complete, then M is complete in the given sub-Riemannian metric.*

Proof. Let $\{x_j\}$ be a Cauchy sequence with respect to d. Then it is a Cauchy sequence with respect to d_R (since $d_R \leqslant d$) and so there exists $x \in M$ such that $x_j \to x$ in the d_R metric. But topologically the two metrics are equivalent, so $x_j \to x$ in the d metric. q.e.d.

For example, suppose $M = \mathbf{R}^n$ and $|g^{jk}(x)| \leqslant c(1 + |x|)$ for all j, k, and all x. Then it is easy to contract to a Riemannian metric satisfying the same estimate, and it is easy to show that the Riemannian metric is complete. Thus M is complete.

8. Isometries

Let M and $\tilde{M}$ be sub-Riemannian manifolds of the same dimension and subdimension, and let $\psi\colon M \to \tilde{M}$ be a homeomorphism.

Definition 8.1. We say that ψ is an *isometry* if ψ preserves distance, $\tilde{d}(\psi(x), \psi(y)) = d(x,y)$ for all $x, y \in M$. We say that ψ is an *infinitesimal isometry* if ψ is C^1 and

$$\tilde{g}(\psi(x)) = d\psi(x)\, g(x)\, d\psi(x)^* \tag{8.1}$$

(we do not assume $d\psi$ is surjective). We say that an infinitesimal isometry is *regular* if

$$\psi(\exp_P u) = \exp_{\psi(P)}\big(d\psi(P)^* u\big) \tag{8.2}$$

for every point $P \in M$ and cotangent $u \in T_P^*$.

Theorem 8.2. (a) *An infinitesimal isometry is an isometry.*

(b) *An infinitesimal isometry of class C^2 is a diffeomorphism and is regular.*

(c) *An isometry of class C^1 is an infinitesimal isometry.*

Proof. (a) Note that (8.1) implies that $d\psi(x)$ maps a subspace of S_x, namely the image of $g(x)\,d\psi(x)^*$ onto $\tilde{S}_{\psi(x)}$, hence the assumption that $\dim S_x = \dim \tilde{S}_{\psi(x)}$ implies that $d\psi(x)$ maps S_x one-to-one onto $\tilde{S}_{\psi(x)}$. Now a routine calculation from (8.1) shows that ψ preserves lengthy curves and their length, hence it preserves distance. Thus ψ is an isometry.

(b) Now it follows that ψ must preserve length minimizing geodesics. Fix a point P and let $x(t)$ be any length minimizing geodesic passing through P at $t = 0$. Let $y(t) = \psi(x(t))$ be the image length minimizing geodesic through $\psi(P)$. If $(y(t), \eta(t))$ is a cotangent lift satisfying (H-J), then a straightforward computation shows that $(x(t), d\psi(x(t))^*\eta(t))$ is a cotangent lift satisfying (H-J) (the computation involves second derivatives of ψ). Setting $t = 0$ we find that $d\psi(P)^*$ maps the cotangents corresponding to length minimizing geodesics at $\psi(P)$ onto the cotangents corresponding to length minimizing geodesics at P. But we have observed in Theorem 6.4 that these are sets of positive measure, so the linear map $d\psi(P)^*$ must be surjective. This shows that ψ is a diffeomorphism. Finally we repeat the above calculations for a general geodesic to establish (8.2).

(c) A C^1 isometry preserves C^1 lengthy curves and their lengths, and by differentiating the length of such a curve with respect to the parameter, we obtain that $d\psi(x)$ must map S_x isometrically onto $S_{\psi(x)}$. By dualizing this statement we obtain (8.1).

Lemma 8.3. (a) *A regular infinitesimal isometry is determined by the values of $\psi(P)$ and $d\psi(P)$ for a fixed P in M.*

(b) *Assume the strong bracket generating hypothesis. If ψ_j is a sequence of regular infinitesimal isometries such that $\psi_j(P) \to Q$ and $d\psi_j(P) \to h$ for some point Q in $\tilde{M}$ and some linear transformation h: $T_P \to T_Q$, then there exists a regular infinitesimal isometry ψ such that $\psi_j \to \psi$ uniformly on compact sets.*

Proof. (a) By (8.2), ψ is determined by $\psi(P)$ and $d\psi(P)$ on all geodesics emanating from P, and by iteration and Corollary 3.3 it is determined everywhere.

(b) By (8.2) we have $\psi_j(\exp_P u) = \exp_{\psi(P_j)}(d\psi_j(P)^*u)$ and the right side clearly has the limit $\exp_Q(h^*u)$ as $j \to \infty$. Thus if we define ψ on the image of $\exp_P$ by $\psi(\exp_P u) = \exp_Q(h^*u)$, then ψ is well defined (if $\exp_P u = \exp_P v$, then $\exp_Q(h^*u) = \exp_Q(h^*v)$) and ψ is the pointwise limit of ψ_j. Furthermore the limit is uniform if u is restricted to a bounded set, and by Corollary 7.2 this shows the convergence of ψ_j to ψ is uniform on a sufficiently small neighborhood of P.

Now the definition of ψ shows that it is C^2 (in fact C^∞) in the image of any neighborhood in which $\exp_P$ is a diffeomorphism. Since ψ is the limit of isometries it is an isometry, so by Theorem 8.2 applied locally it must be a regular infinitesimal isometry on such open sets. If Q is a point in such a set, then by applying (8.2) at Q we see that ψ is C^2 in the image of any neighborhood on which $\exp_Q$ is a diffeomorphism. But by the results of §5 we can reach any point of M in a finite number of steps, and so ψ is C^2 everywhere, hence a regular infinitesimal isometry. It is also clear from the definition of ψ that the derivatives of ψ_j converge to derivatives of ψ on such neighborhoods, and so by iteration we obtain the convergence of ψ_j to ψ uniformly on compact sets.

Lemma 8.4. (a) *The set of isometries of M to M with fixed point P forms a compact group with the topology of uniform convergence of compact sets, called the isotropy group of P.*

(b) *Assume the strong bracket generating hypothesis. Then the subgroup of the isotropy group of P of regular infinitesimal isometries is closed, hence compact.*

Proof. (a) This is a straightforward application of the Arzela-Ascoli theorem, since the isometric property gives the uniform equicontinuity estimate.

(b) Let ψ_j be a sequence of regular infinitesimal isometries with fixed point P, and consider the linear transformations $d\psi_j(P)^*$ on T_P^*. Let B_ε denote the subset of P unit cotangents $u \in T_P^*$ such that $\exp_P(tu)$ is a length minimizing geodesic on $[0, \varepsilon]$. By Corollary 7.2, B_ε is compact for ε small enough. Clearly each $d\psi_j(P)^*$ must preserve B_ε. It is clear that B_ε contains a basis of T_P^*, so $d\psi_j(P)^*$ must lie in a compact subset of the space of linear transformations on T_P^*, hence by passing to a subsequence we can make $d\psi_j(P)^*$ converge, and hence ψ_j converges to a regular infinitesimal isometry by the previous lemma.

Theorem 8.5. *Assume the strong bracket generating hypothesis. Then the set of regular infinitesimal isometries of M to M with the compact-open topology is a Lie group G, and the subgroup G_P of those isometries with fixed point P, is a compact subgroup isomorphic to a subgroup of the isometries of S_P. In particular, a regular infinitesimal isometry is determined by $\psi(P)$ and $d\psi(P)$ restricted to S_P.*

Proof. Consider the set H of all regular infinitesimal isometries with fixed point P such that $d\psi(P)$ is the identity on S_P. By the previous lemmas H is a compact Lie group and $\psi \to d\psi(P)$ is an isomorphism of H onto a group H' of linear transformations of T_P. Our goal is to show that H' consists of the identity alone.

It is convenient to think in terms of a matrix representation of H'. We choose a basis of T_P so that the first m elements span S_P. Then the matrices of

elements of H' have the form

$$d\psi(P) = \begin{pmatrix} I & B \\ 0 & D \end{pmatrix}.$$

In terms of a dual basis for T_P^*, where the last $n - m$ elements span N_P, we have

$$d\varphi(P)^* = \begin{pmatrix} I & 0 \\ B^* & D^* \end{pmatrix}.$$

To show $D = I$ we use the raised Christoffel symbol Γ: $T_P^*/N_P \times N_P \to S_P$. For $\xi \in T_P^*$ and $v \in N_P$ we have the transformation law

$$d\psi(P)\Gamma\big(d\psi(P)^*\xi, d\psi(P)^*v\big) = \Gamma(\xi, v)$$

by Lemma 2.3 and the fact that ψ is a regular infinitesimal isometry (this argument uses the fact that ψ is C^2, which was established in the course of the proof of Lemma 8.3). Now $\Gamma \in S_P$ so $d\psi(P)\Gamma = \Gamma$ and $d\psi(P)^*v = D^*v$, so we have $\Gamma(d\psi(P)^*\xi, D^*v) = \Gamma(\xi, v)$. But also $d\psi(P)^*\xi$ differs from ξ by a null cotangent, hence by Lemma 2.3 we have $\Gamma(\xi, D^*v) = \Gamma(\xi, v)$. Then by Theorem 2.4 $\Gamma(\xi, \cdot)$ is injective if ξ is nonnull, so we have $D^*v = v$ hence $D = I$.

Thus we have shown that H' is a subgroup of the group of matrices of the form $\left(\begin{smallmatrix} I & B \\ 0 & I \end{smallmatrix}\right)$, which is isomorphic to a Euclidean space. But H' is compact, and the only compact subgroup of Euclidean space is the identity.

Now if ψ_1 and ψ_2 are in G_P with $d\psi_1(P) = d\psi_2(P)$ on S_P, then $\psi_1 \circ \psi_2^{-1}$ is the identity by the above argument. This shows $\psi \to d\psi(P)|_{S_P}$ is an isomorphism of G_P with a subgroup of the isometries of S_P.

Finally, the fact that G forms a Lie group follows from a general theorem of Montgomery and Zippin [31, pp. 208 and 212], to the effect that if G is any locally compact effective transformation group on a C^1 manifold with each transformation G of class C^1, then G is a Lie group and the action $G \times M \to M$ is C^1. To apply this theorem in our case we need to verify that G is locally compact. The essential argument has already been given in Lemma 8.3(b). To apply this we need to show that as ψ varies over all elements of G that map P to a compact set K, the linear transformations $d\psi(P)$ are confined to a compact set. But we have verified this if K is the set $\{P\}$, and it follows easily if K is a singleton set $\{Q\}$. But the set of $d\psi(P)$ varies continuously with Q, and so another compactness argument completes the verification.

Definition 8.6. A *Killing vector field* is a vector field Y satisfying $\nabla_{\text{sym}} Y \equiv 0$, where ∇_{sym} is given by Definition 2.5.

Theorem 8.7. *The C^2 Killing vector fields form a Lie algebra, and the subalgebra of complete C^2 Killing vector fields is naturally isomorphic to the Lie algebra of G (under the strong bracket generating hypothesis). If the manifold is complete, then all C^2 Killing vector fields are complete.*

Proof. Let ψ_t be a one-parameter family of C^2 mappings of M to M (i.e., $\psi_t \circ \psi_s = \psi_{t+s}$) depending differentiably on t, and let Y be the derivative vector field ($Y(\psi_t) = \dot{\psi}_t$). Then a simple computation shows that ψ_t is an infinitesimal isometry for all t if and only if Y is a Killing vector field. If Y is C^2, then ψ_t is C^2 hence regular by Lemma 8.2(b), while if ψ_t is regular, then ψ_t is C^∞ under the strong bracket generating hypothesis. Clearly Y is complete if and only if ψ_t is globally defined. Now suppose the manifold is complete; if Y is a C^2 Killing field, then for small enough t_0 there exists a nonempty open set U such that if $x \in U$, then $\psi_t(x)$ is defined for $0 \leqslant t \leqslant t_0$ by $\dot{\psi}_t(x) = Y(\psi_t(x))$ and $\psi_0(x) = x$, and ψ_t is a local isometry. But the completeness then shows that U is also closed, and so ψ_t is globally defined.

A direct computation shows that the bracket of two C^2 Killing vectors is also a Killing vector field, and the above arguments applied locally show that a C^2 Killing vector field is automatically C^∞, under the strong bracket generating hypothesis. q.e.d.

A sub-Riemannian manifold is called *homogeneous* if it has a transitive group of regular infinitesimal isometries. It is easy to see that such a manifold is automatically complete, since every point has a complete closed ball around it, and the radius may be taken independent of the point.

9. Sub-Riemannian symmetric spaces

Definition 9.1. A *sub-Riemannian symmetric space* is a sub-Riemannian manifold M which has a transitive Lie group G of regular infinitesimal isometries acting differentiably on M with the following properties:

(i) The isotropy subgroup K of a point P is compact.

(ii) K contains an element ψ such that $d\psi(P)|_{S_P} = -I$ and $\psi^2 = I$.

It is easy to see that if these properties hold at one point, then they hold at every point. If we assume the strong bracket generating hypothesis we can dispense with (i) and the condition $\psi^2 = I$, since these are consequences of the other hypotheses.

If G is a group for which (i) and (ii) hold, we will call G an *admissible isometry group* for M. For a given M there may be more than one admissible isometry group.

Theorem 9.2. *If M is a sub-Riemannian symmetric space and G is an admissible isometry group, then there exists an involution σ of G such that $\sigma(K) \subseteq K$ with the following properties (we write $\mathfrak{g} = \mathfrak{g}^+ + \mathfrak{g}^-$, where $\mathfrak{g}^\pm$ are the subspaces of $\mathfrak{g}$ on which $d\sigma$ acts as $\pm I$):*

(a) *$\mathfrak{g}$ is generated as a Lie algebra by a subspace $\mathfrak{p}$ and the subalgebra $\mathfrak{k}$ with $\mathfrak{p} \subseteq \mathfrak{g}^-$, $\mathfrak{k} \subseteq \mathfrak{g}^+$, where $\mathfrak{k}$ is the Lie algebra of K.*

(b) *There exists a positive definite quadratic form Q on $\mathfrak{p}$ and $\operatorname{ad} K$ maps $\mathfrak{p}$ to itself and preserves Q. Furthermore, $\mathfrak{p}$ may be identified with S_P under the exponential map (of the Lie algebra $\mathfrak{g}$), and Q with the sub-Riemannian metric on S_P.*

Conversely, given a Lie group G and an involution σ such that (a) and (b) hold, then G/K forms a sub-Riemannian symmetric space, where $S_P = \exp \mathfrak{p}$ for the point P identified with the coset K, and the sub-Riemannian metric on S_P is given by Q. The bundle S and its metric is then uniquely determined by the requirement that elements of G be infinitesimal isometries.

Proof. Given a sub-Riemannian symmetric space choose an admissible isometry group G and a point P, let K be the isotropy subgroup of P, and $\mathfrak{k}$ its Lie algebra. We identify M with G/K and T_P with $\mathfrak{g}/\mathfrak{k}$. We define $\sigma(h) = \psi \circ h \circ \psi$ so that $d\sigma = \operatorname{Ad} \psi$, where ψ is the elements of K given in (ii). Clearly σ is an involution of G because $\psi^2 = I$ and $\sigma(K) \subseteq K$.

Now we consider the adjoint action of K on $\mathfrak{g}$. For any $k \in K$, $\operatorname{Ad} k$ factors to a linear map on $\mathfrak{g}/\mathfrak{k}$, and because K is compact we can find a complementary space $\mathfrak{p}_1$ preserved by $\operatorname{Ad} K$. We can identify $\mathfrak{p}_1$ with the tangent space at P under the exponential map, and define $\mathfrak{p} \subseteq \mathfrak{p}_1$ to be the inverse image of S_P. Then $\operatorname{Ad} K$ on $\mathfrak{p}_1$ under the identification is equal to dk on T_P. The condition $d\psi(P)|_{S_P} = -I$ implies that $\mathfrak{p} \subseteq \mathfrak{g}^-$. It is easy to see that the condition that S be bracket generating is equivalent to the condition that $\mathfrak{k}$ and $\mathfrak{p}$ generate $\mathfrak{g}$. We define the quadratic form Q on $\mathfrak{p}$ by taking the metric Q_P on S_P under the identification $S_P = \exp \mathfrak{p}$, and then $\operatorname{Ad} K$ preserves $\mathfrak{p}$ and Q since the elements of K are infinitesimal isometries. Thus we have verified all of (a) and (b) except the condition $\mathfrak{k} \subseteq \mathfrak{g}^+$.

Now if $k \in K$ and $\operatorname{Ad} k$ is equal to the identity on $\mathfrak{p}$, it follows that $\operatorname{Ad} k$ is equal to the identity on $\mathfrak{p}_1$ since $\mathfrak{p}$ generates it, and hence $dk = I$ at P hence k is the identity. In particular, if $X \in \mathfrak{k}$ and X commutes with $\mathfrak{p}$, then $X = 0$. Now if $X \in \mathfrak{k}$ and $Y \in \mathfrak{p}$, then $[X - d\sigma X, Y] \in \mathfrak{p}$ so

$$-[X - d\sigma X, Y] = d\sigma[X - d\sigma X, Y] = [d\sigma X - X, d\sigma Y] = [X - d\sigma X, Y],$$

and so $X - d\sigma X$ commutes with $\mathfrak{p}$. Thus we have $X - d\sigma X = 0$ hence $\mathfrak{k} \subseteq \mathfrak{g}^+$.

Conversely, given G and σ, we define a sub-Riemannian metric on G/K as follows. For P the point identified with the coset K, we define $S_P = \exp \mathfrak{p}$ and $g(P)$: $T_P^* \to S_P$ by $Q(Y, g(P)\xi) = \langle Y, \xi \rangle$ for all $Y \in S_P$ as in (2.1). Given

any $x \in G/K$ find h in G such that $h(P) = x$ and set $g(x) = dh(P)g(P)dh(P)^*$. It follows from (b) that the definition does not depend on the choice of h, and each element of G is an infinitesimal isometry. Since G/K has a real analytical structure it follows from Theorem 8.2(b) that these are regular infinitesimal isometries.

It remains to construct the isometry ψ called for in (ii). We define $\psi(hP) = \sigma(h)P$ for any $h \in G$, which is unambiguous because $\sigma(K) \subseteq K$. Now a simple computation involving the chain rule and the inverse function theorem shows that ψ is an infinitesimal isometry, and ψ is real-analytic hence regular. Clearly $d\psi(P) = -I$ on $\mathfrak{p}$ because $\mathfrak{p} \subseteq \mathfrak{g}^-$ and $\psi^2 = I$ because $\sigma^2 = I$. If $\psi \in K$, we are done. If $\psi \notin K$, then we enlarge G by adjoining ψ. We obtain the disjoint union of G and $G\psi$ because $\psi \circ h \circ \psi = \sigma(h)$, and so the new K is still compact. q.e.d.

As a corollary of the proof we note that any isometry in K is determined by the restriction of its derivative at P to S_P. Unfortunately, it is not clear in general whether the full group of regular infinitesimal isometries is admissible, since we do not have a proof of the compactness of the isotropy group of a point without the strong bracket generating hypothesis. Any admissible group has dimension at most $n + m(m-1)/2$.

We could also consider giving the data for a sub-Riemannian symmetric space entirely in infinitesimal form. For this we would take a Lie algebra $\mathfrak{g}$, an involution of $\tilde{\sigma}$ of $\mathfrak{g}$ with $\mathfrak{g}^{\pm}$ its eigenspaces corresponding to ± 1, $\mathfrak{k} \subseteq \mathfrak{g}^+$ a subalgebra and $\mathfrak{p} \subseteq \mathfrak{g}$ a subspace, and a positive definite quadratic form Q on $\mathfrak{p}$, such that $\operatorname{ad} \mathfrak{k}$ preserves $\mathfrak{p}$ and Q. The theorem shows how to associate this data to a sub-Riemannian symmetric space. Conversely, if we take G to be the simply-connected and connected Lie group with Lie algebra $\mathfrak{g}$, we can always lift $\tilde{\sigma}$ to an involution σ of G such that $d\sigma = \tilde{\sigma}$. Let $K = \exp \mathfrak{k}$. We need to assume that K is compact. Then it is easy to see that G/K is a sub-Riemannian symmetric space with the given data.

It is important to note that the involution ψ is not the geodesic symmetry at the point P, since we will not have $d\psi(P) = -I$ on the whole tangent space.

If φ is any automorphism of G such that $\varphi(K) \subseteq K$ and $d\varphi$ preserves $\mathfrak{p}$ and Q, then φ factors to a regular infinitesimal isometry of G/K preserving P by the same reasoning as in the proof of the theorem. In general we cannot expect to obtain the full isotropy group of P in this way.

We conclude this section with some examples:

(1) Let G be any connected noncompact semisimple Lie group, K the identity subgroup (*not* a maximal compact subgroup), σ a Cartan involution (with respect to a maximal compact subgroup K_1), and $\mathfrak{p} = \mathfrak{g}^-$. It is not hard to show that $\mathfrak{p}$ generates $\mathfrak{g}$ (if $\mathfrak{g}$ is simple this follows from the fact that

$\mathfrak{p} + [\mathfrak{p}\mathfrak{p}]$ is an ideal, and in general $\mathfrak{g}$ splits into a direct sum of simple ideals). We could take any positive definite quadratic form on $\mathfrak{p}$, but a natural choice would be the restriction of the Killing form. In that case the group K_1 acts as isometries preserving the identity. In particular, if we take the Lorentz group $SO_e(n, 1)$, then we obtain the maximal dimension for the isometry group. Although p is always a two-step generator in these examples, the strong bracket generating hypothesis is usually not satisfied.

(2) Let G be a compact semisimple Lie group, K the identity subgroup, σ any nontrivial involution, $\mathfrak{p} = \mathfrak{g}^-$, and Q the negative of the Killing form. Then as before $\mathfrak{g} = \mathfrak{p} + [\mathfrak{p}\mathfrak{p}]$, and the subgroup K_1 corresponding to $\mathfrak{g}+$ acts as an isotropy group of the identity. The example of the rotation group $SO(n + 1)$ with σ equal to conjugation by the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & I_n \end{smallmatrix}\right)$ has an isometry group of maximal dimension.

(3) Let G be the free 2-step nilpotent Lie group on m generators ($2 \leqslant m$). Then $\mathfrak{g}$ has a basis X_j and Y_{jk} for $1 \leqslant j \leqslant m$ and $1 \leqslant j < k \leqslant m$ with $[X_j X_k] = Y_{jk}$ for $j < k$ and Y_{jk} in the center, and G is the simply-connected Lie group with Lie algebra $\mathfrak{g}$. We take K as the identity, and σ defined by $d\sigma(X_j) = -X_j$, $d\sigma(Y_{jk}) = Y_{jk}$ so that $\mathfrak{p}$ is spanned by the X_j. For Q we take the form that makes $\{X_j\}$ an orthonormal basis. Then again the entire orthogonal group of $(\mathfrak{p}, Q)$ extends to automorphisms of $\mathfrak{g}$ which lift to G, hence the isometry group has maximal dimension. But again the strong bracket generating hypothesis fails except for $m = 2$.

(4) Let G be the n-dimensional Heisenberg group and K the identity group. A basis for $\mathfrak{g}$ is X_j, Y_j, Z with $1 \leqslant j \leqslant n$, and $[X_j, Y_j] = 2Z$ with all other brackets zero. We define σ by $\sigma(X_j) = -X_j$, $\sigma(Y_j) = -Y_j$, $\sigma(Z) = Z$ so that $\mathfrak{p}$ is spanned by the X_j and Y_j, and we choose Q to make $\{X_j\} \cup \{Y_j\}$ an orthonormal basis. The strong bracket generating hypothesis holds for this example. Not every orthogonal transformation on $\mathfrak{p}$ extends to an automorphism of $\mathfrak{g}$, but those that do can be identified with the unitary group $U(n)$ (essentially by considering $X_j + iY_j$ as a complex variable), so the isotropy subgroup is at least transitive on the unit sphere of $\mathfrak{p}$. This example has been studied extensively by a number of authors independently ([3], [25], [32]) and the (H-J) equations for geodesics can be solved explicitly. There is also a group of dilations $X_j \to \lambda X_j$, $Y_j \to \lambda Y_j$, $Z \to \lambda^2 Z$ for $\lambda > 0$.

10. Three-dimensional symmetric spaces

In this section we give a classification up to local isometry of all sub-Riemannian symmetric spaces in three dimensions. Of course when $n = 3$ we automatically have the strong bracket generating hypothesis.

We begin by studying the infinitesimal data $\mathfrak{g}$, $\tilde{\sigma}$, $\mathfrak{k}$, $\mathfrak{p}$, Q. We must have $\dim \mathfrak{p} = 2$ and $\dim \mathfrak{g} = 3$ or 4, with $\mathfrak{g} = \mathfrak{p} + [\mathfrak{p}\mathfrak{p}] + \mathfrak{k}$ and so $\mathfrak{g}^+ = [\mathfrak{p}\mathfrak{p}] + \mathfrak{k}$, $\mathfrak{g}^- = \mathfrak{p}$. It is then easy to see from the Jacobi identity that $\mathfrak{p} + [\mathfrak{p}\mathfrak{p}]$ is a subalgebra. Thus we can initially assume $\dim \mathfrak{g} = 3$ and $\mathfrak{k} = 0$, and then decide whether or not we can adjoin a one-dimensional $\mathfrak{k}$.

Let X_1, X_2 be a basis for $\mathfrak{p}$, and let $Y = [X_1, X_2]$. Then $\operatorname{ad} Y$ must preserve $\mathfrak{g}^-$, and the Lie algebra structure is determined entirely by specifying a 2×2 real matrix A such that $[YX_1] = a_{11}X_1 + a_{12}X_2$ and $[YX_2] = a_{21}X_1 + a_{22}X_2$. The Jacobi identity is equivalent to $\operatorname{trace} A = 0$. If we take a different basis for $\mathfrak{p}$ given by $\tilde{X} = MX$ for a nonsingular real 2×2 matrix M, then A is transformed to $(\det M)MAM^{-1}$. It is then a simple exercise in linear algebra to show that up to isomorphism there are six distinct possibilities for $\mathfrak{g}$ and $\mathfrak{p}$:

(1) $A = 0$, in which case $\mathfrak{g}$ is the Heisenberg Lie algebra, $[X_1X_2] = Y$, $[YX_1] = [YX_2] = 0$. Here $\mathfrak{g}$ is nilpotent.

(2) $A = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, in which case $\mathfrak{g}$ is the Lie algebgra of the motion group of the Euclidean plane, $[X_1X_2] = Y$, $[YX_1] = X_2$, $[YX_2] = 0$. Here X_2 and Y span an abelian subalgebra corresponding to translations of the plane, and X_1 corresponds to rotations, and $\mathfrak{g}$ is solvable.

(3) $A = \left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$, in which case $\mathfrak{g}$ is the Lie algebra of the motion group of the Lorentzian plane, $[X_1X_2] = Y$, $[YX_1] = -X_2$, and $[YX_2] = 0$. Here X_2 and Y span an abelian subalgebra corresponding to translations, and X_1 corresponds to Lorentz transformations, and $\mathfrak{g}$ is solvable.

(4) $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, in which case $\mathfrak{g} = so(3)$, $[X_1X_2] = Y$, $[YX_1] = X_2$, $[X_2Y] = X_1$. Here $\mathfrak{g}$ is compact semisimple.

(5) $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, in which case $\mathfrak{g} = sl(2, \mathbf{R})$, $[X_1X_2] = Y$, $[YX_1] = X_2$, $[YX_2] = X_1$. Here $\mathfrak{g}$ is noncompact semisimple.

(6) $A = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, in which case $\mathfrak{g} = sl(2, \mathbf{R})$, $[X_1X_2] = Y$, $[YX_1] = -X_2$, $[YX_2] = X_1$. Here $\mathfrak{g}$ is noncompact semisimple.

The difference between case (5) and case (6) is in the involution, which is equal to the usual Cartan involution in case (6) but not in case (5).

Now we specify the positive definite quadratic form on $\mathfrak{p}$ by a matrix $Q = \left(\begin{smallmatrix} a & b \\ b & d \end{smallmatrix}\right)$ with $a > 0$, $d > 0$, $ad > b^2$. If M is any nonsingular matrix such that $(\det M)MAM^{-1} = A$, then the basis $\tilde{X} = MX$ is equivalent and the corresponding matrix $\tilde{Q}$ satisfies $M^{\mathrm{tr}}\tilde{Q}M = Q$. Thus to complete the description of the infinitesimal data we need to specify one Q in each equivalence class. We summarize the results:

(1): any nonsingular M is allowed, so there is only one equivalence class, and we can take $Q = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

(2) and (3): any $M = \left(\begin{smallmatrix} \pm 1 & \lambda \\ 0 & \mu \end{smallmatrix}\right)$ with $\mu \neq 0$ is allowed, so there is a one-

parameter family of equivalence classes, and $Q = \left(\begin{smallmatrix} a & 0 \\ 0 & 1 \end{smallmatrix}\right)$ with $a > 0$ gives a representative of each class.

(4) and (6): any orthogonal M is allowed, so there is a two-parameter family of equivalence classes, and $Q = \left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$ with $a \geqslant d > 0$ gives a representative of each class.

(5) any Lorentzian M is allowed, so there is a two-parameter family of equivalence classes, and $Q = \left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$ with $a > 0$, $d > 0$ gives a representation of each class.

Next we construct sub-Riemannian symmetric spaces corresponding to the six classes of data. We can always take $M = G$ where G is the simply-connected Lie group with Lie algebra $\mathfrak{g}$, but this is not always the most transparent choice. Instead we prefer a unified treatment in which G is a subgroup of $\mathrm{SL}(3, \mathbf{R})$ and the involution on G is conjugation with the matrix

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and G is the image under the exponential map of $\mathrm{SL}(3, \mathbf{R})$ of a Lie algebra isomorphic to $\mathfrak{g}$.

(1) G is the Heisenberg group of matrices

$$\begin{pmatrix} 1 & 0 & x_2 \\ x_1 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathfrak{g} = \left\{ \begin{pmatrix} 0 & 0 & x_2 \\ x_1 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} \right\}.$$

(2) G is the proper Euclidean motion group of matrices

$$\begin{pmatrix} \cos x_1 & \sin x_1 & x_2 \\ -\sin x_1 & \cos x_1 & y \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathfrak{g} = \left\{ \begin{pmatrix} 0 & x_1 & x_2 \\ -x_1 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} \right\}.$$

(3) G is the proper orthochronous Poincaré group of matrices

$$\begin{pmatrix} \cosh x_1 & \sinh x_1 & x_2 \\ \sinh x_1 & \cosh x_1 & y \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathfrak{g} = \left\{ \begin{pmatrix} 0 & x_1 & x_2 \\ x_1 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} \right\}.$$

(4) G is the rotation group $SO(3)$ and

$$\mathfrak{g} = \left\{ \begin{pmatrix} 0 & x_1 & x_2 \\ -x_1 & 0 & -y \\ -x_2 & y & 0 \end{pmatrix} \right\}.$$

(5) G is the proper orthochronous Lorentz group $SO_e(2, 1)$ and

$$\mathfrak{g} = \left\{ \begin{pmatrix} 0 & x_1 & x_2 \\ -x_1 & 0 & -y \\ x_2 & -y & 0 \end{pmatrix} \right\}.$$

(6) $G = SO_e(1,2)$ and

$$\mathfrak{g} = \left\{ \begin{pmatrix} 0 & x_1 & x_2 \\ x_1 & 0 & y \\ x_2 & -y & 0 \end{pmatrix} \right\}.$$

Next we compute the full isometry group for each of these spaces. It suffices to find all regular infinitesimal isometries that preserve the identity, for these together with G generate all regular infinitesimal isometries. To do this we need to find all orthogonal transformations of $\mathfrak{p}$ (with respect to Q) which extend to automorphisms of $\mathfrak{g}$. If we identify the transformation with the 2×2 matrix M we require $M^{\mathrm{tr}}QM = Q$ for M to be orthogonal and $(\det M)MAM^{-1} = A$ for M to extend to an automorphism (via $Y \to (\det M)Y$) of $\mathfrak{g}$.

Now the four element group of transformations

$$\begin{pmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{pmatrix}$$

for ε_1 and ε_2 taking values ± 1 is easily seen to satisfy these conditions. On the other hand, in order to have a larger group we must be in one of three cases: (1), (4) with $a = d$, or (6) with $a = d$. In these cases the full orthogonal group $O(2)$ will satisfy both conditions. In cases (4) and (6) with $a = d$ the associated isometries of G are given simply by conjugation with $\left(\begin{smallmatrix} 1 & 0 \\ 0 & M \end{smallmatrix}\right)$ for $M \in O(2)$. In case (1) it is a little more complicated to describe the isometries in terms of the coordinates (x_1, x_2, y) of the point

$$\begin{pmatrix} 1 & 0 & x_2 \\ x_1 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}.$$

Here the group law is

$$(x_1, x_2, y) \circ (x_1', x_2', y') = (x_1 + x_1', x_2 + x_2', y + y' + x_1 x_2').$$

Corresponding to the improper matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ the isometry is $(x_1, x_2, y) \to (x_1, -x_2, -y)$. Corresponding to the proper rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

the isometry is

$$(x_1, x_2, y) \to (x_1 \cos\theta - x_2 \sin\theta,\ x_1 \sin\theta + x_2 \cos\theta,\ y - x_1 x_2 \sin^2\theta + \tfrac{1}{2}(x_1^2 - x_2^2) \sin\theta\cos\theta).$$

We note that in case (1) there is an additional group of dilations, namely $(x_1, x_2, y) \to (\lambda x_1 \cdot \lambda x_2 \cdot \lambda^2 y)$ for $\lambda > 0$. Each dilation multiplies lengths by the fixed value λ. None of the other examples possess nonisometric dilations.

All in all, there is good reason to think of case (1) as the analogue of flat Euclidean space, case (4) with $a = d$ as the analogue of the spheres of constant positive curvature, and case (6) with $a = d$ as the analogue of hyperbolic spaces of constant negative curvature. It is tempting to interpret the expression $-[[XY]Z]$ for $X, Y, Z \in \mathfrak{p}$ as an analogue of the curvature tensor in the Riemannian case, as this would be consistent with the above analogies. However, in general this expression will not have the expected symmetries of the Riemannian curvature tensor.

11. Local geometry

Riemannian geometry is locally Euclidean, with curvature serving as a measure of the higher order deviation from Euclidean. Sub-Riemannian geometry, on the other hand, has a more complicated local behavior. As a simple example consider the Heisenberg group geometry (Example 1 of §10). The existence of dilations shows that small neighborhood are similar to large neighborhoods, so nothing is gained by working locally; in particular there are no approximately Euclidean triangles.

In this section, we will answer some simple questions about distances, triangles, and cut points. We choose a point P and a local coordinate system with P as the origin such that the tangents to the coordinate directions $x^1, \cdots, x^m$ form an orthonormal basis for S_p at the origin, hence $g^{jk}(0) = \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$. We write $x = (x', x'')$ for x in the coordinate patch with $x' \in \mathbf{R}^m$ and $x'' \in \mathbf{R}^{n-m}$, and we use the conventions of §5 to denote indices restricted to $1, \cdots, m$ by Roman letters a, b, etc., and indices restricted to $m+1, \cdots, n$ by Greek letters α, β, etc.

Theorem 11.1. *Assume S is a two-step bracket generator. There exists a neighborhood of the origin on which the distance to the origin $d(x, 0)$ is bounded above and below by a constant multiple of $|x'| + |x''|^{1/2}$. In particular, for any Riemannian contraction with metric d_R, the estimate $d(x, y) \leqslant c d_R(x, y)^{1/2}$ holds on any compact set.*

Proof. To show $d(x, 0) \leqslant c(|x'| + |x''|^{1/2})$ we need to construct a lengthy curve from 0 to x of length at most $c(|x'| + |x''|^{1/2})$. To do this we choose a basis $X_1, \cdots, X_m$ for the bundle S in a neighborhood of the origin such that $X_a^j(0) = \delta_a^j$. By the hypothesis that S is a 2-step bracket generator we can find

$n - m$ pairs of indices $a(\alpha)$, $b(\alpha)$ such that if we set $Y_\alpha = [X_{a(\alpha)} X_{b(\alpha)}]$, then $X_1, \cdots, X_m, Y_{m+1}, \cdots, Y_n$ form a basis for the tangent bundle in a neighborhood of the origin.

Let $\varphi_a(t)$ denote the local flow generated by X_a. The orbits of φ_a will be lengthy curves, as will be all curves that are piecewise orbits. We denote one of these piecewise orbits by $\Phi(P, (a_1, t_1), (a_2, t_2), \cdots, (a_k, t_k))$, meaning the curve starts at P, then follows φ_{a_1} for the time t_1, then φ_{a_2} for the time t_2, and so on. If any t_j is negative we follow $\varphi_{a_j}(-t)$ for the time interval $-t_j$. Now it is well known that the curve

$$t \to \begin{cases} \Phi\big(P, (a_2, -\sqrt{t}), (a_1, -\sqrt{t}), (a_2, \sqrt{t}), (a_1, \sqrt{t})\big), & t > 0, \\ \Phi\big(P(a_1, -\sqrt{|t|}), (a_2, -\sqrt{|t|})(a_1, \sqrt{|t|}), (a_2, \sqrt{|t|})\big), & t < 0, \end{cases}$$

has derivative $[X_{a_1}, X_{a_2}]$. Given variables $t_1, \cdots, t_n$ in a neighborhood of the origin, we consider the lengthy curve

$$\begin{aligned} &\gamma(t_1, \cdots, t_n) = \Phi(0, (1, t_1), (2, t_2), \cdots, (m, t_m), \\ &\big(b(m+1), -\sqrt{t_{m+1}}\big), \big(a(m+1), -\sqrt{t_{m+1}}\big), \\ &\big(b(m+1), \sqrt{t_{m+1}}\big), \big(a(m+1), \sqrt{t_{m+1}}\big), \cdots, \\ &\big(b(n), -\sqrt{t_n}\big), \big(a(n), -\sqrt{t_n}\big), \big(b(a), \sqrt{t_n}\big), \big(a(a), \sqrt{t_n}\big)\big) \end{aligned}$$

when all $t_\alpha \geqslant 0$ (with the obvious modification if some $t_\alpha < 0$). This curve has length bounded by a multiple of $|t'| + |t''|^{1/2}$. If we consider the endpoint of the curve, call it x, then the map $t \to x$ is C^1 and has an invertible derivative at the origin of the form $\left(\begin{smallmatrix} I & 0 \\ * & * \end{smallmatrix}\right)$, so by the inverse function theorem there is a neighborhood of the origin on which the inverse map $x \to t$ yields the lengthy curve $\gamma(t)$ which joins 0 to x with length bounded by a multiple of $|x'| + |x''|^{1/2}$. This shows $d(0, x) \leqslant c d_R(0, x)^{1/2}$, and we obtain the bound $d(x, y) \leqslant c d_R(x, y)^{1/2}$ uniformly on compact sets by a routine argument.

For the bound from below we need to show $|x''|^{1/2} \leqslant c d(x, 0)$, since we always have $d_R(x, 0) \leqslant d(x, 0)$. Let $x(t)$ be any lengthy curve with $x(0) = 0$ parametrized by arc length. Then $d_R(x(t), 0) \leqslant d(x(t), 0) \leqslant t$ so $|g(x(t)) - g(0)| \leqslant ct$. Now we can write $\dot{x}(t) = g(x(t))\xi(t)$ for a cotangent lift $\xi(t)$ which can be chosen so that $|\xi(t)| \leqslant c$. Thus

$$|\dot{x}^\alpha(t)| = |g^{\alpha p}(x(t))\xi_p(t)| \leqslant ct$$

since $g^{\alpha p}(0) = 0$ hence we obtain $|x''(t)| \leqslant ct^2$. By choosing a length minimizing geodesic from 0 to x we have $|x''| \leqslant c d(0, x)^2$ as desired. q.e.d.

To generalize this result to higher brackets it is necessary to use the Campbell-Baker-Hausdorff-Dynkin formula. A particularly clear exposition

can be found in [33]. The result is also given in [30]. A further development of these ideas is given in [39].

Next we consider triangles. From a fixed point P we follow two geodesics to $\exp_P tu$ and $\exp_P t\tilde{u}$ and then join the endpoints by a length minimizing geodesic, where $gu \neq 0$ and $g\tilde{u} \neq 0$ and t is small enough that the two initial geodesics are length minimizing. The behavior of the length of the third geodesic, $d(\exp_P tu, \exp_P t\tilde{u})$, as $t \to 0$, can be specified at least as to order of magnitude.

Theorem 11.2. *Assume S is a two-step bracket generator. If $gu \neq g\tilde{u}$, then there exist constants c_1 and c_2 such that*

$$c_1 t \leqslant d(\exp_P(tu), \exp_P(t\tilde{u})) \leqslant c_2 t.$$

If $gu = g\tilde{u}$ then

$$d(\exp_P(tu), \exp_P(t\tilde{u})) \leqslant ct^{3/2}.$$

Proof. The upper bound $c_2 t$ is obvious, since we can always join $\exp_P(tu)$ to $\exp_P(t\tilde{u})$ by a broken geodesic through P. Now in §5 we computed the Taylor expansion

$$\exp_P(tu)^k = tg^{kp}(P)u_p - \frac{t^2}{2}\Gamma^{kp_1p_2}(P)u_{p_1}u_{p_2} + O(t^3).$$

Thus if $gu \neq g\tilde{u}$ we have $d_R(\exp_P(tu), \exp_P(t\tilde{u})) \geqslant c_1 t$ which establishes the lower bound.

Now suppose $gu = g\tilde{u}$, so that $\tilde{u} = u + v$ with $gv = 0$. Then

$$\exp_P(t\tilde{u})^k = tg^{kp}(P)u_p - \frac{t^2}{2}\Gamma^{kp_1p_2}(P)u_{p_1}u_{p_2} - t^2\Gamma^{kp_1p_2}u_{p_1}v_{p_2} + O(t^3)$$

since $\Gamma^{kp_1p_2}v_{p_1}v_{p_2} = 0$. Now by Lemma 2.3 we have $\Gamma^{kp_1p_2}u_{p_1}v_{p_2} = g^{kj}w_j$ for some w. We consider the point $\exp_{(\exp_P(tu))}(-t^2w)$, which is clearly at distance $O(t^2)$ from $\exp_P(tu)$. To complete the proof we need to show that the distance to $\exp_P(t\tilde{u})$ is $O(t^{3/2})$. But the Taylor expansion (with coordinates centered at P) is

$$\exp_{(\exp_P(tu))}(-t^2w)^k = \exp_P(tu)^k - t^2g^{kp}(\exp_P(tu))w_p + O(t^4)$$

and so we have

$$\exp_{(\exp_P(tu))}(-t^2w) = \exp_P(t\tilde{u}) + O(t^3)$$

since $g^{kp}(\exp_P(tu)) = g^{kp}(P) + O(t)$. Thus the Riemannian distance between the points is at most $O(t^3)$ and the result follows from the previous Theorem. q.e.d.

Remark. In the case of the Heisenberg group it is possible to compute the distance exactly. If $gu \neq g\tilde{u}$, then

$$\lim_{t\to 0} t^{-1} d(\exp_P(tu), \exp_P(t\tilde{u}))$$

exists, while if $gu \neq g\tilde{u}$ then

$$\lim_{t\to 0} t^{-3/2} d(\exp_P(tu), \exp_P(t\tilde{u}))$$

exists and is nonzero. It would be interesting to know if these limits exist more generally. Distinct geodesics with $gu = g\tilde{u}$ form what might be called "horn angles" at P.

Next we consider the question of uniqueness of length minimizing geodesics. The *cut locus* of P is defined to be the set of all points such that there exist more than one length minimizing geodesic joining the point to P, and any point of the cut locus is called a *cut point*. We have already observed that a geodesic from P cannot be length minimizing beyond the first cut point. In contrast to the case of Riemannian geometry, cut points occur arbitrarily close to P.

Theorem 11.3. *Assume the strong bracket generating hypothesis. Then for every sufficiently small ε there exist at least one cut point of distance ε from P.*

Proof. Choose ε small enough that the closed ball of radius ε about P is complete (hence compact). Suppose that there were no cut points on the sphere $S_\varepsilon(P)$ of radius ε. Look at the corresponding set E in T_P^*; i.e., E is the set of u such that $\langle gu, u\rangle = \varepsilon^2$ and $\exp_P(tu)$ for $0 \leqslant t \leqslant 1$ is length minimizing. Then the map $u \to \exp_P(u)$ would be a one-to-one continuous map of E onto $S_\varepsilon(P)$. Now we have seen that under the strong bracket generating hypothesis the set E is compact, hence the map would be a homeomorphism. But it is impossible for a compact subset of the cylinder $\langle gu, u\rangle = \varepsilon^2$ (topologically $S^{m-1} \times \mathbf{R}^{n-m}$) to be homeomorphic to $S_\varepsilon(P)$ which is the boundary of an open set in $\mathbf{R}^n$.

Remark. If $m = n - 1$, then we can show that there are at least two cut points on each sphere. By the results of §5 we know there must be a collar $S^{m-1} \times [-\varepsilon_0, \varepsilon_0]$ contained in E, and we can argue separately that each of the two components of the complement of the collar gives rise to at least one cut point.

12. Analysis of sub-Laplacians

Let M be an n-dimensional manifold and let $X_1, \cdots, X_m$ be m linearly independent real vector fields which are bracket generating. We call $L = \sum_{j=1} X_j^2 + Y$ a *sub-Laplacian*, where Y is any real vector field. These operators

were first studied by Hormander (without the hypothesis of linear independence) who showed they are hypoelliptic, and have since been the subject of intense study (e.g. [36]).

Associated to L there is a sub-Riemannian metric. In terms of the tangent bundle, we take S to be the span of $X_1, \cdots, X_m$ and we choose the quadratic form at each point that makes $X_1(x), \cdots, X_m(x)$ an orthogonal basis for S. In terms of the metric $g^{jk}(x)$, we simply express L in local coordinates and set

$$L = g^{jk}(x)\frac{\partial}{\partial x^j}\frac{\partial}{\partial x^k} + \text{first order terms.}$$

If $X_p = a_p^j(x)\partial/\partial x^j$, then

$$g^{jk} = \sum_{p=1}^{m} a_p^j(x)a_p^k(x).$$

Now suppose we are given a smooth density, $d\mu = G(x)\,dx$ in local coordinates, for G a smooth positive function. We adjust the first order term in L to make L formally symmetric with respect to $L^2(d\mu)$. Let X_j^* denote the formal adjoint of X_j with respect to the inner product $\langle f_1, f_2\rangle = \int_M f_1(x)f_2(x)G(x)\,dx$. Then $X_p^* = -X_p - \operatorname{div} a_p - G^{-1}X_pG$ and so we will have $L = -\sum_{p=1}^m X_p^*X_p$ provided we take $Y = \sum_{p=1}^m c_pX_p$ with

$$c_p = \operatorname{div} a_p + G^{-1}X_pG. \tag{12.1}$$

From now on we will make this choice of Y.

We could, conversely, start with Y and ask if there is a density which gives rise to it. Clearly Y must be a linear combination of the X_p, but there will be other compatibility requirements. However, if there is such a density, it is unique up to a constant multiple since (12.1) determines $X_p \log G$ and we have

Lemma 12.1. *If h is any function such that $X_ph = 0$ for all p, then h is constant.*

Proof. If all $X_ph = 0$, then clearly $[X_p, X_q]h = 0$ and similarly for all higher brackets. Thus $\nabla h = 0$ so h is constant. q.e.d.

Now suppose the manifold M is complete in the sub-Riemannian metric associated to L (note that this condition is independent of the density). Then essentially all the results of [38] concerning the Laplacian are valid for L, with essentially the same proofs. The key point is Theorem 7.3 which is the exact analogue of Lemma 2.2 of [38]. We will not repeat the proofs of [38] here, but we give a sampling of some of the results:

(1) L is essentially self-adjoint on the domain $G_{\text{com}}^{\infty}(M)$;

(2) The heat semigroup $e^{t\Delta}$ is contractive on all $L^p(d\mu)$, $1 \leqslant p \leqslant \infty$, and $e^{t\Delta}f$ gives the unique solution to the Cauchy problem $u_t = Lu$, $u(0) = f$ under the hypothesis that $\|u(\cdot, t)\|L_{p(d\mu)}$ is uniformly bounded for $1 < p < \infty$.

(3) The axioms of the Littlewood-Paley-Stein theory of [10] are satisfied, so for example $(-L)^{is}$ and $(I-L)^{is}$ are bounded operators on L^p for $1 < p < \infty$;

(4) The heat operator $e^{t\Delta}$ is given by a positive C^∞ kernel.

References

[1] J. Baillieul, *Geometric methods for nonlinear optimal control problems*, J. Optim. Theory Appl. **25** (1978) 519–548.

[2] J.-M. Bismut, *Large deviations and the Malliavin Calculus*, Progress in Math., Vol. 45, Birkhäuser, Basel, 1984.

[3] R. W. Brockett, *Control theory and singular Riemannian geometry*, New Directions in Applied Mathematics (P. J. Hilton and G. S. Young, eds.), Springer, Berlin, 1981, 11–27.

[4] _____, *Nonlinear control theory and differential geometry*, Proc. Internat. Congr. Math., Warsaw, 1983, 1357–1368.

[5] C. Carathéodory, *Untersuchungen über die Grundlagen der Thermodynamik*, Math. Ann. **67** (1909) 355–386.

[6] _____, *Calculus of variations and partial differential equations of the first order*, Holden-Day, San Francisco, 1965.

[7] L. Cesari, *Optimization-theory and applications*, Springer, Berlin, 1983.

[8] S. S. Chern & J. Moser, *Real hypersurfaces in complex manifolds*, Acta Math. **133** (1974) 219–271.

[9] W. L. Chow, *Uber Systeme Von Lineaaren Partiellen Differentialgleichungen erster Orduung*, Math. Ann. **117** (1939) 98–105.

[10] M. Cowling, *Harmonic analysis on semi-groups*, Ann. of Math. (2) **117** (1983) 267–283.

[11] C. Fefferman, *Monge-Ampere equations, The Bergman kernel and geometry of pseudo-convex domains*, Ann. of Math. (2) **103** (1967) 395–416; *Erratum* **104** (1976) 393–394.

[12] B. Franchi & E. Lanconelli, *Une metrique associe a une classe d'operateurs elliptiques degeneres*, Proc. Linear, Partial, and Pseudo-Differential Operators, Rend. Sem. Math. Univ. E. Polytech., Torino, 1982.

[13] _____, *Une condition géométrique pour l'inégalité de Harnack*, J. Math. Pures Appl. **64** (1985) 237–256.

[14] B. Gaveau, *Principle de moindre action, propogation de la challeur et estimates sous-elliptiques sur certains grouppes nilpotent*, Acta Math. **139** (1977) 95–153.

[15] _____, *Systemes dynamiques associe a certains operateurs hypoelliptiques*, Bull. Sci. Math. **102** (1978) 203–229.

[16] M. Gromov, *Structures metriques pour les varietes Riemanniennes*, Cedic, Paris, 1981.

[17] N. C. Günther, *Hamiltonian mechanics and optimal control*, Thesis, Harvard University, 1982.

[18] U. Hamenstädt, *On the geometry of Carnot-Carathéodory metrics*, preprint, 1986.

[19] R. Hermann, *Some differential-geometric aspects of the Lagrange variational problem*, Illinois J. Math. **6** (1962) 634–673.

[20] _____, *The differential geometry of foliations*. II, J. Math. Mech. **11** (1962) 303–315.

[21] _____, *Geodesics of singular Riemannian metrics*, Bull. Amer. Math. Soc. **79** (1973) 780–782.

[22] _____, *Differential geometry and the calculus of variations*, 2nd ed., Math. Sci. Press, Brookline, MA, 1977.

[23] D. Jerison & J. M. Lee, *A subelliptic, nonlinear eigenvalue problem and scalar curvature on CR manifolds*, Contemporary Math., Vol. 27, Amer. Math. Soc., Providence, RI, 1984, 57–64.

[24] W. Klingenberg, *Riemannian geometry*, de Gruyter, Berlin, 1982.

[25] A. Koranyi, *Geometric aspects of analysis on the Heisenberg group*, Topics in Modern Harmonic Analysis, Vol. II, Proc. Sem. (Torino and Milano), May-June 1982, 209–258.

[26] _____, *Geometric properties of Heisenberg-type groups*, Advances in Math. **56** (1985) 28–38.

[27] A. Koranyi & H. M. Riemann, *Quasiconformal mappings in the Heisenberg group*, Invent. Math. **80** (1985) 309–338.

[28] _____, *Horizontal normal vectors and conformal capacity of spherical rings in the Heisenberg group*, preprint.

[29] R. Léandre, *Integration dans la fibre associée à une diffusion dégénérée*, preprint.

[30] J. Mitchell, *On Carnot-Carathéodory metrics*, J. Differential Geometry **21** (1985) 35–45.

[31] D. Montgomery & L. Zippin, *Topological transformation groups*, Interscience, New York, 1955.

[32] A. I. Nachman, *The wave equation on the Heisenberg group*, Comm. Partial Differential Equations **7** (1982) 675–714.

[33] A. Nagel, E. M. Stein & S. Wainger, *Balls and metrics defined by vector fields*. I, Acta Math. **155** (1985) 103–147.

[34] P. Pansu, *An isoperimetric inequality on the Heisenberg group*, Differential Geometry on Homogeneous Spaces, Rend. Sem. Mat. Torino, Fascicolo Speciale, 1983, 159–174.

[35] _____, *Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rank* 1, preprint.

[36] L. P. Rothschild & E. M. Stein, *Hypoelliptic differential operators and nilpotent groups*, Acta Math. **137** (1976) 247–320.

[37] A. Sanchez-Calle, *Fundamental solutions and geometry of the sum of squares of vector fields*, Invent. Math. **78** (1984) 143–160.

[38] R. S. Strichartz, *Analysis of the Laplacian on a complete Riemannian manifold*, J. Funct. Anal. **52** (1983) 48–79.

[39] _____, *The Campbell-Baker-Hausdorff-Dynkin formula and solutions of differential equations*, J. Funct. Anal., to appear.

[40] H. Sussmann, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc. **180** (1973) 171–188.

[41] T. J. S. Taylor, *Some aspects of differential geometry associated with hypoelliptic second order operators*, preprint.

[42] _____, *Off diagonal asymptotics of hypoelliptic diffusion equations and singular Riemannian geometry*, preprint.

[43] S. M. Webster, *Pseudohermetian structures on a real hypersurface*, J. Differential Geometry **13** (1978) 25–41.

Cornell University